CHEROKEE NATION
System Solutions

APHIS Big Data Recommendations

October 19, 2017

Submitted by:
Cherokee Nation System Solutions, LLC
Trayci Koppie, Technical Project Manager
C: (703) 898-5942 | email: Trayci.Koppie@cnt-fc.com
777 W. Cherokee St. Catoosa, OK 74015 |

4803 Innovation Drive, Suite 3 Fort Collins, CO 80525|

This document includes data that shall not be disclosed outside the Government and shall not be duplicated, used, or disclosed—in whole or in part—for any purpose other than to evaluate its contents. If, however, a contract or delivery order is awarded to this offeror as a result of—or in connection with—the submission of this data, the Government shall have the right to duplicate, use, or disclose the data to the extent provided in the resulting contract. This restriction does not limit the Government's right to use information contained in this data if it is obtained from another source without restriction. The data subject to this restriction is contained in all sheets marked with the following legend: "Use or disclosure of data contained on this sheet is subject to the restriction on the title page of this proposal."

Table of Contents

1 Executive Summary

Project Overview

The United States Department of Agriculture (USDA) Animal and Plant Health Inspection Service (APHIS) is in need of overarching and scalable IT solutions for scientific computing environments to ensure that now, and in the future labs and scientists have access to the appropriate IT resources and tools, can efficiently exchange data among colleagues for research purposes, and invaluable research data is secure and backed up to prevent loss or damage. Cherokee Nation System Solutions (CNSS) was contracted by APHIS to assess scientific computing in the Riverdale, MD, Beltsville, MD, Ames, IA and Fort Collins, CO facilities and to review the Veterinary Services (VS), Science, Technology and Analysis Services (STAS) assessment by BioTeam. Based on our assessments and analysis, this report provides CNSS' findings, identified challenges and gap analysis, and our short and long-term strategy recommendations to assist APHIS in moving forward with their Big Data initiatives.

Summary of Findings (SOF)

CNSS Summary of Findings (SOF) (*provided as Appendix B*), is based on interviews conducted at Ames, IA, Ft. Collins, CO, Beltsville and Riverdale, MD and numerous field sites via teleconference and WebEx with representatives from APHIS leadership, division directors, scientists, lab-techs, and IT staff focused on reviews of relevant documentation, and in-depth, on-premises assessments of the relevant IT computing, data storage, and networking equipment.

The following points comprise the most salient SOF and represent the kinds of factors considered as foremost opportunities and challenges:

- Some sites are experimenting with the ARS Scientific Converged Infrastructure Network (SCINet), Amazon Web Services Cloud (AWS), and Microsoft Azure Cloud.
- In general, APHIS scientists find scientific computing is no longer well served by local or enterprise computing resources as storage is limited, resources do not scale well to manage surges, outlying offices have major connection challenges, and these limitations impede timely processing, collaborative research, and tool sharing among departments.
- Network connectivity is hampering numerous scientific endeavors for multiple programs at several of the sites.
- The growth of scientific data is outpacing the ability for Enterprise IT to keep up on its legacy storage systems, in some programs the growth trend is over 200% annually.
- Data storage collection, consolidation, management, access, and sharing suffer from multiple challenges of storage and communication capacity, timeliness, redundancy, security, and related problems.
- Communication and storage issues ranging from moving data around digitally in elaborate workaround schemes to time-consuming portable storage device shipping are impeding Sharing / Collaboration.
- Confidentiality, integrity, and availability are at risk on numerous, unsupported portable storage devices, and ad hoc computing environments utilizing repurposed workstations that are sometimes disconnected from the enterprise IT environment to avoid update induced processing interruption. In addition, these devices are not subject to APHIS Continuity of Operations (COOP) and pose security risks.
- Lack of dedicated scientific IT support staff to maintain the scientific IT environment and to assist Scientists with effective utilization of the scientific IT environment, software optimization, and development of scientific tools and workflows.

To mitigate the challenges scientists continue to find new and novel ways to reclaim IT assets or cleverly relocate data to get around the lack of available storage. They need a more effective, dedicated Scientific Information Technology (SIT) environment that provides sufficient capacity, performance, information assurance, security, resiliency, COOP, and disaster recovery capabilities to support modern scientific workflows.

Gaps and Recommendations

Based on the analysis in the SOF and available technology, CNSS recommends that APHIS continue to use and expand its partnership with Agriculture Research Service (ARS) and SCINet for their SIT environment.

CNSS recommends that APHIS continue to develop the Microsoft Azure Platform as their Enterprise Information Technology infrastructure empowering their regulatory and governance focused tasks. Amazon Web Services (AWS) is more conducive to scientific computing given that APHIS can take advantage of the modular design that enables "serverless" computing, allowing scientists to concentrate on building models for their research and not infrastructure. SCINet currently takes advantage of AWS with their hybrid-cloud to surge computing power and services on-demand as needed. APHIS can take advantage of AWS's platform to make cloud-based infrastructure available on-demand lowering Capital Expenditures (CAPEX) and allowing scientists to use computing power only when needed. These hybrid initiatives empower APHIS scientists and satisfy DCOI requirements.

CNSS provides a high-level summary below of the near and long-term solutions that APHIS could leverage to get to the desired state where scientific data and processing are segregated from IT assets to optimize the performance of both systems.

CNSS Summary of Gaps and Recommendations		
GAP	**Recommendation**	**Pros/Cons**
High-Performance Computing (HPC) – need centralized or cloud-based platforms that scale processing, provide data management dynamically, can be accessed/shared easily, minimize CAPEX investments, and is in line with DCOI.	**Leverage ARS SCINet** to all extents possible, including adding SCINet points of presence (POP) to PPQ building 580 and Riverdale. **Leverage ARS SCINet Connectivity** to move large datasets between APHIS and adopted AWS solutions. **Leverage Azure Stack** as an on-premises resource to augment the Microsoft hosted Azure solution (hybrid cloud) currently implemented by APHIS until full implementation through SCINet or AWS. Azure Stack is recommended for Enterprise data vs. scientific data. Since Azure Stack charges only for resources consumed, it is more cost-effective than buying dedicated servers.	**SCINet** does not contain a POP at Bldg. 580 or Riverdale, MD. **Azure Stack** can be installed by the vendor at no initial cost, and reoccurring costs are based on resources consumed. **Azure Stack integrates seamlessly** with Microsoft hosted Azure instances. **Provides centralized control** over environment both on-premises and in the Azure cloud
Big data and metadata storage and management - scientific data needs to be centralized to reduce the burdens of the Enterprise Storage platform, minimize data duplication across the networks, and improve accessibility.	**SCINet is the long-term goal** with some short-term realization in Ames, IA and Ft. Collins, CO. **Replace aging Oracle Pillar Axioms with Nimble C5K Storage Arrays.** **Consolidate mobile storage** solutions into temporary storage systems, such as WD My Cloud. **Implement File Integrity Monitoring (FIM)** and metadata options to ensure data integrity and accessibility **iRods is a free and open-source product** that meets APHIS meta-data indexing needs.	**ARS SCINet** does not contain a secure enclave that is accredited to contain APHIS data. **Centralization of data** allows for a site with lackluster connectivity to take advantage by employing Virtual Desktop Infrastructure (VDI). **FIM allows APHIS** to ensure the data integrity at rest in archives and ARS SCINet and reduces archival management through manual processes.
High-speed communication	**ARS SCINet affords the internal** and external communications bandwidth and a	**Lack of APHIS/ARS policies** on

CNSS Summary of Gaps and Recommendations		
GAP	**Recommendation**	**Pros/Cons**
for data sharing and collaboration – Sharing research data, software code and experimental methods are the backbone of gathering the range of observations, confirming scientific results, and translating research to speed up discoveries and identify large-scale trends.	collaborative ecosystem that will enable APHIS to expand its global plant and animal health leadership in collaboration with external agencies and educational institution partners. **Roll out VDI to the program sites** outside of Ames, IA, Ft. Collins, CO and Riverdale, MD. **Leverage VDI** for sites that have low-speed communications to work on their data locally. **WD My Cloud** PR4100 as a temporary on-premises solution until data can be lifted from the site and stored in ARS SCINet, AWS or Enterprise storage. AWS Snowball **would allow data** destined for AWS or even ARS SCINet to be lifted physically in 7 days from sites with low-speed connectivity and large datasets. **Establish an enterprise** code repository	how data will be shared, stored, and priority of processing on Ceres. **Transport of scientific data** to AWS could occur via AWS Snowball while to centralize APHIS scientific data and subsequently into SCINet. **Local vault based storage** is not optimal at remote sites, but necessary to consolidate the data for long-term storage strategies
Effective Scientific IT support resources – to manage the SIT environment and optimize Scientists' use of it	**Cultivate a parallel capacity** to the Virtual Research Support Core (VRSC) to provide guidance and support to APHIS Scientists	**With APHIS and ARS VRSC IT** staff working together, skill gaps and solutions can be more rapidly formed to meet each program need in SCINet
Permanent big data leadership – that provides Scientific IT policies, an acquisition strategy, and a detailed plan for optimizing, consolidating, and migrating systems.	**Create SIT branch** parallel to current Enterprise IT organization. **Organization-wide artifacts** must ensure both buy-in and common understanding **Coordination and integration** of diverse functions	**ARS has created a Chief Scientist Information Office (CSIO)** to great effect. APHIS lacks central scientific direction to help guide IT in creating meaningful agency-wide programs/infrastructure.
Program Level Requirements – specialized program level requirements for a comprehensive big data strategy.	**Recommended program level** strategies include VDI, workstation upgrades, HPCs, FileMaker Pro, and numerous near-term options unique to each program need.	**Program-sponsored solutions** are separate from APHIS IT initiatives and do not have the support or funding for enterprise IT. **Near-term solutions** that meet the immediate and unique needs of the program while APHIS centralizes its big data initiative which could take years.
Access to Open Source Tools – Lack of access has proven to be a challenge for APHIS Enterprise IT and a long, frustrating process for	**A SIT environment**, collaborating with ARS to move scientific research to SCINet, adoption of AWS and AWS Direct Connect, and providing localized solutions as needed, will provide Scientists with	**Ability to create and share** their models publicly or privately using open source tools and code **Adaptive, robust environment**

CNSS Summary of Gaps and Recommendations		
GAP	**Recommendation**	**Pros/Cons**
Scientists.	access to numerous open source tools.	that allows APHIS to take advantage of these tools as they evolve rapidly. **Some programs have access** to tools due to legacy grandfathering while others do not have equal access.

These recommended solutions can help APHIS and their programs realize a simpler, more integrated environment where tools for APHIS program personnel are made available to empower their work and scientific research.

Rough Order of Magnitude (ROM)

CNSS compared the pricing for 1 PB of storage for both Azure and AWS and recommends that a combination of both best serve APHIS data needs with AWS focusing on scientific purposes while the Azure data would be more enterprise in nature.

AWS snowball is being recommended as a solution that will allow for the transport of large datasets in the organization to an AWS data repository. Data transport could be done collectively as a single data lift into AWS or at the program level where large amounts of data would be moved using the device. Microsoft offers a competitive device called Azure Data Box recently released as a community preview in September 2017. There is not enough information from Microsoft to gather whether there are enough units to lift 1 PB of data or what the current pricing model will be. We recommend that APHIS wait until Microsoft has made this product more mature and instead utilize Microsoft's data ingest service or the 1 Gbps connection between APHIS and Azure Cloud.

AWS and Azure have been compared for HPC to contrast what the costs would be in both environments. While its most likely that APHIS will use AWS for Scientific computing and Azure for Enterprise computing, both solutions are capable of doing HPC. AWS is the best choice for scientific computing as its leveraged in ARS SCINet.

VDI was compared in Azure Cloud along with an on-premises solution that is currently implemented in Ames, IA supporting 150 desktop users. APHIS should continue to grow its VDI solutions in Ft. Collins, CO, and Riverdale, MD to allow remote users to access their datasets as they are consolidated at these locations. Azure Cloud VDI was added as a comparison to an on-premises installation and is also relevant when applications or large datasets reside in Azure that require a local high-speed connection. This approach would allow the user to access the datasets and manipulate them as they would be collocated with the software and data. The following shows the ROM for the CNSS recommendations discussed in detail in this document.

Section	Name	Quantity	Subtotal
Storage	AWS 1 Petabyte (PB) Storage Estimate	1	$30,242/mo
HPC Cluster	AWS HPC Cluster – Cloud Only	4	$22,413.31/mo + one -time $16,924.00
Snowball – Data migration services	AWS Snowball Data Migration services	13	$3,250
Storage	Azure 1 PB Storage Estimate	1	$44,409.85/mo
Azure HPC Cluster	Azure HPC Cluster – Cloud Only	8	$20,197.12/mo
Azure VDI	Azure VDI – Cloud Only	~3000	~$31,800/mo
On-premises VMWare VDI	VMWare VDI – On-premises	~3000	~$32,400/mo
		Total One Time: Total Annual:	$20,174.00 $2,177,547.36

1.1 Purpose & Context

The purpose of this report is to propose both near and long-term strategies to move forward with Big Data initiatives as defined by APHIS AG-32KW-P-17-0089 contract SOW CLIN 0001 – Onsite Big Data Assessment. We base Recommendations on the As-Is APHIS technology environment report that CNSS delivered, Summary of Findings (SOF) on September 11, 2017.

The CNSS contract CLIN 0001 targets an assessment on agency Big Data activities report that captures the current state of activities and a strategy to move forward with Big Data initiatives within the agency. Activities include:

- **Interviews with APHIS Big Data Working Group to capture requirements** - As noted, SOF is based on interviews CNSS conducted at Ames, IA, Ft. Collins, CO, Beltsville, MD and Riverdale, MD and numerous field sites via teleconference with representatives from APHIS leadership, division directors, Scientists, lab-techs, and IT staff focused on reviews of relevant documentation, and in-depth, on-premises assessments of the relevant IT computing, data storage, and networking equipment.
- **Documentation of the baseline for current Big Data activities –** SOF provides a comprehensive assessment of the Big Data activities baseline and is referenced throughout this report.
- **Documentation of future strategies to implement a Big Data approach for APHIS** – Please see 2 Options Analysis & Recommendations and 3 Summary of Recommendations sections of this report for both near and long-term strategies
- **Analysis of the following options –** We present in-depth analysis details for each of the following in 2.3 Long-Term Strategies of this report:
    - **Program-sponsored solutions -** Big data computing solutions would be developed at the program level without an agency-wide sponsorship, support or resources.
    - **Consolidated APHIS/ARS cloud2 -** APHIS would collaborate with ARS scientific big data efforts, including implementation of a hybrid cloud, with both on-premises and cloud-based infrastructure. The cloud would be used for extremely intensive data analysis, backups, and disaster recovery. (Note: This is the alternative that ARS chose).
    - **APHIS-only hybrid cloud -** Develop a hybrid cloud (on-premises and cloud-based infrastructure) with no sharing resources with ARS.
    - **APHIS-only public cloud -** Other than network-related connectivity, there would be no on-premises infrastructure. All computing and storage resources would be located in the cloud, including production and alternate backup/disaster recovery sites.

In addition to the site visits and interviews, CNSS also reviewed all of the documents delineated in Appendix A: Documents Reviewed.

1.2 Summary of Findings

The following points comprise the most salient SOF and represent the kinds of factors considered as foremost opportunities and challenges.

- **The perceived current state of scientific computing as served by enterprise and localized IT resources** is that limited access to sufficient scientific technology processing, data storage, analysis, integration, security, sharing, and transfer of technical resources impede timely processing, collaborative research, and tool sharing.
    - Scientific computing is no longer well served by local or enterprise computing resources as storage is limited, it does not scale well to manage surges, and outlying offices have major connection challenges.

  - The growth of scientific data is outpacing the ability for Enterprise IT to keep up on its legacy storage systems, in some programs the growth trend is over 200% annually.
  - Data storage collection, consolidation, management, access, and sharing suffer from multiple challenges of storage and communication capacity, timeliness, redundancy, security, and related problems.
  - Sharing / Collaboration are impeded by communication and storage issues ranging from moving data around digitally in elaborate workaround schemes to shipping portable storage devices – both of which are time-consuming.
  - Confidentiality, integrity, and availability are at risk on numerous, unsupported portable storage devices, and ad hoc computing environments utilizing repurposed workstations that are sometimes disconnected from the enterprise IT environment to avoid update induced processing interruptions.

- **What Scientists are doing** today can be characterized as an innovative collaboration with IT resources to work within the current local and enterprise IT computing environment. To mitigate the challenges that programs in APHIS have faced, they continue to find new and novel ways to reclaim IT assets or cleverly relocate data to get around the lack of available storage. The following are examples of what is being done and to what effect.
  - Sites are experimenting with ARS Scientific Converged Infrastructure Network (SCINet), Amazon Web Services Cloud (AWS), and APHIS has a substantial investment in Microsoft Azure Cloud. Some have upgraded local computing resources, and VS is using the ARS Ceres HPC Cluster to some benefit. As noted in SOF, these efforts represent both stopgap and long-term potential strategies.

Getting access to open source suites or creating repositories for code has been a painful process for VS in Ames, IA and other APHIS *programs due to the lead time required gaining approval to allow even an evaluation of the software.*

  - Use of uncontrolled storage devices such as memory sticks, hard drives, and other network shares to store and ship data enables Scientists marginally. Often these systems aren't part of the Enterprise Disaster Recovery Strategy so reliance could lead to catastrophic loss of invaluable scientific data and leaves the organization open to malware infection.
  - Most sites have local computational resources in the form of workstations and other reclaimed IT assets. Some locations are unable to power all of their workstations in the ad-hoc labs due to building power and cooling constraints.
  - Some Scientists have found that ad-hoc environments impede the timely processing of data and have turned to IT for the more centralized computational resources. IT has provided centralized servers in VS and PPQ running Red Hat Enterprise Linux. Most of the Scientists do not have experience in a command line environment and are more comfortable in a Graphical User Interface (GUI) environment leading to lack of adoption of the provided platform.

- **APHIS IT Infrastructure Big Data Baseline –** The following are examples of the significant information presented in SOF.
  - Centralized Infrastructure
    - APHIS has a dedicated connection to the Azure environment through the USDA Multiprotocol Label Switching (MPLS) Cloud (UTN) and an APHIS 1 Gigabyte (GB) connection that directly connects the Ft. Collins Site to Azure. Applications targeted to move are mostly enterprise in nature.

VS in Ames, IA contains an active 100 GB connection to Internet2 segregated from APHIS *Enterprise Network that allows fast transport between the Amazon Web Services (AWS) cut out into the ARS SCINet Network*

        - APHIS's SCINet connection is a benefit of ARS and APHIS sharing resources in collocated facilities.
        - The VS Bioinfo drive is the shared storage that NVSL/CVB uses to converge the PC/MAC environment with the Red Hat Enterprise Linux Servers on APHIS Enterprise Infrastructure (AEI). This environment is not currently replicated between Ames, IA and Ft. Collins, CO via the APHIS Enterprise Network.
        - There is an evident lack of storage available in the enterprise storage system which is growing at an average rate of 235% per year.
    - Non-Centralized Infrastructure
        - VS has three High-End Linux Servers - one in Plum Island, NY and two in Ames, IA. PPQ located in Building 580 in Beltsville, MD is getting one similar to VS.
        - VS has open source tools that PPQ labs are unable to take advantage of due to security constraints preventing their installation and use on the APHIS corporate network.
        - Scientists in Ames are the predominant users of the Ceres systems with a few in Ft. Collins, but slow data transfer from APHIS to Ceres hampers the initiative.
        - As VS IT has time to install a workstation in the Café (the lunch room) and in various labs at Ames, IA, NVSL facility, scientists will be able to transfer data from portable or network devices.
        - End of Life unsupported and frequently unknown to IT, repurposed workstations are used by a variety of sites, including East and West.
        - The network connectivity and bandwidth to access the proposed centralized and cloud platforms are not sufficient, which results in WS field sites being unable to leverage their remote computational assets.
        - DCOI has the greatest impact on non-centralized assets and compounds the already challenging field office communication issues.
        - CPHST and PGQP labs do not have any centralized or dedicated computational platform or any form of dedicated hardware for computational processing or storage on site. Instead, a majority of the work is done on local workstations or decommissioned hardware that is air-gapped from the network to run custom scripts for day-to-day tasks. CPHST and PGQP Labs are the largest producers of Big Data in the PPQ Program. Working with Next Generation Sequencers, they are on target to produce 30TB of raw and analytical data per year.
    - Continuity of Operations (COOP)
        - COOP is inconsistent outside of the enterprise environment since data is stored on unsupported devices which are outside of IT policy and Security Policy and are not factored into the Enterprise COOP plan. The issue is compounded by misperceptions of who is backing up the data.
        - Current data size constraints prevent using the VS Bioinfo drive as an offsite backup and COOP site for Ames, IA Scientific Data. Ames data is backed up to tape and shipped to Ft. Collins, CO. Full replication of Ames scientific data is planned when the storage solution is modernized.

- **Current and Future Needs** – Scientists are increasingly dependent on a more effective, dedicated Scientific IT (SIT) environment that provides sufficient capacity, performance, information assurance, security, resiliency, continuity of operations, and disaster recovery capabilities to support modern scientific workflows. A widely-held perception is that APHIS needs a dedicated SIT environment separate from IT, operating with a less restrictive security posture oriented toward the performance needed for scientific workflows. Such an environment comprises:

> *AWS modular services for high throughput processing to Peta Scale Storage and Machine Learning Algorithms would give Scientists a large modular platform that can be used to meet the emerging needs.*

- **High-Performance Computing (HPC)** – Scientific computing consumes vast amounts of computation, communication, and data management resources and these requirements drive the trend toward cloud-based platforms that scale processing and data management dynamically, can be accessed and shared easily, minimize CAPEX investments, and is in line with DCOI.
- **Big data and metadata storage and management** - Taking into account past and present trends, all sites surveyed are on target to need a dedicated SIT environment to meet data growth projected to be in the Petabytes. It is increasingly evident that scientific data needs to be centralized to reduce the burdens of the Enterprise Storage platform, minimize data duplication across the networks, and improve accessibility.
- **High-speed communication for data sharing and collaboration** – Sharing research data, software code and experimental methods are the backbone of gathering the range of observations, confirming scientific results, and translating research to speed up discoveries and identify large-scale trends.
- **Effective Scientific IT support resources –** to manage the SIT environment and optimize Scientists' use of it, optimize software, and develop scientific tools and workflows. APHIS IT will need to understand each unique programmatic requirement that allows them to maintain their focus on scientific research. By creating a working group at each program level, Enterprise IT will have a better understanding of the unique IT needs and serve more effectively as a collaborative partner in future solutions.
- **Permanent big data leadership –** that provides Scientific IT policies, an acquisition strategy, and a detailed plan for optimizing, consolidating, and migrating systems.
- **Program Level Requirements –** Although enterprise level strategies will provide a solid framework, specialized program level requirements will also need to be considered and addressed to implement a comprehensive big data strategy. Data management for remote sites, licensing and access to specialized tools, unique security constraints such as Confidential Information Protection and Statistical Efficiency Act (CIPSA), and related requirements will remain important considerations for both near and long-term strategies.
- **Access to Open Source Tools –** One of the most powerful collaboration tools that Scientists have today is the ability to create and share their models publically or privately using open source tools and code. To flourish, Scientists need an adaptive, robust environment that allows APHIS to take advantage of these tools as they evolve rapidly. Lack of access has proven to be a challenge for APHIS Enterprise IT and a long, frustrating process for Scientists.

1.3 Options Analysis & Recommendations

Starting with our Analytical Criteria, CNSS then acknowledges DCOI as an important driver for APHIS Scientific and Big Data. CNSS then separates near from long-term challenges and recommendations.

2.2 Near-Term Strategies considers both centralized and non-centralized matters where opportunities exist to begin making a difference now. For the most part, CNSS makes suggestions that will continue to serve APHIS as contributors to long-term solutions. Near-Term Strategies include summaries of challenges and proposed mitigation/resolution strategies related to:

- Microsoft Azure & Azure Stack
- Office Connectivity Challenges for Centralized and Cloud-Based Services
- Data Storage
- The Geospatial Information Systems (GIS)

2.3 Long-Term Strategies convey a programmatic perspective of both enterprise and localized challenges and recommendations including:

- 2.3.1 Program-Sponsored Solutions
- 2.3.2 Consolidated APHIS/ARS Cloud
- 2.3.3 APHIS-only Public & Hybrid Cloud

1.4 Report Organization

In addition to the preceding Overview, this report establishes our analytical criteria, discusses technical and functional aspects, documents how CNSS validated recommendations with vendors and offers Recommendations for moving forward with near and long-term strategies. 4 Rough Order of Magnitude (ROM) conveys high-level cost estimates for each proposed strategy. Appendix A: Documents Reviewed lists documents that were reviewed during the assessment phase, and Appendix B: APHIS Summary of Findings 09/11/17 is a copy of the CNSS SOF document.

2 Options Analysis & Recommendations

2.1 Analytical Criteria

CNSS has conducted in-depth interviews with all APHIS programs in an attempt to ascertain agency-wide, current Big Data needs. During this process, we evaluated:

- Organizations
- People
- Processes
- Infrastructure

During this evaluation, CNSS established an AS-IS baseline of the big data environment in SOF to categorize and plan for the expansion of Big Data in APHIS.

Recommendations are made on the following criteria:

- Leveraging existing and planned infrastructures like SCINet to support Big Data
- Collocating with Agencies like ARS who have a substantial investment in Big Data where there is mutual benefit
- Cloud Technologies
- Innovative Software and Solutions that APHIS is not yet utilizing

Given that APHIS is a large agency with very specific needs at the program level; CNSS has prioritized recommendations to encompass all programs agency-wide. Further recommendations will be made for solutions or software that will integrate into the overall larger infrastructure recommendations.

Datacenter Optimization Initiative (DCOI)

As noted in SOF, one of the largest factors pushing consolidation is DCOI. APHIS is now looking to consolidate a majority of the server technology in offices with less than 20 personnel to Ft. Collins, CO, Riverdale, MD or the Microsoft Azure Cloud.

As APHIS consolidates servers at various locations, the storage and compute capabilities could be moved into:

- Organic Computing at Ft. Collins, CO or Riverdale, MD Datacenters
- Azure Stack at the Ft. Collins, CO or Riverdale, MD Datacenters
- Azure Cloud off-premises

With an overall reduction of 68% of APHIS *compute environment and the elimination of 213TB of organic storage on these sites,* APHIS *needs to either consolidate on centralized Enterprise Storage or move Scientific Data into the cloud or SCINet.*

2.2 Near-Term Strategies

This section concentrates on short-term opportunities that exist to meet big data needs with Scientific and Hybrid Cloud Computing. Topics include the DCOI implementation, Microsoft Azure, connectivity challenges for centralized and cloud-based services, data storage, and Geospatial Information Systems (GIS). The following table ties four of the identified issues and opportunities to evaluated and proposed near-term strategies at a very high level. Following sections delve into the specifics of each strategy.

Near-Term Strategies	Azure	VDI	SCINet	Storage
High-Performance Computing	✓		✓	
Big data and metadata storage and management	✓	✓	✓	✓
High-speed communication for data sharing and collaboration	✓	✓	✓	✓

Near-Term Strategies	Azure	VDI	SCINet	Storage
Program Level Requirements	✓	✓	✓	✓

Microsoft Azure

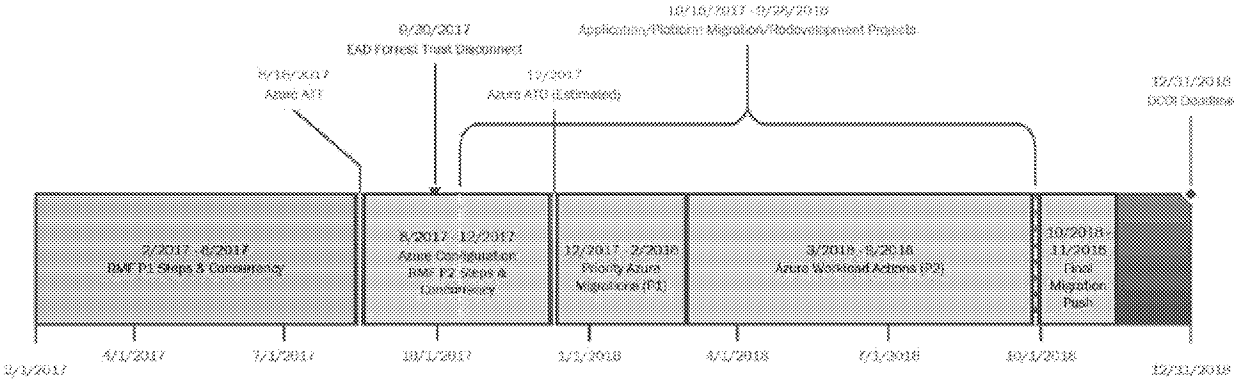
To grow an Azure presence, APHIS could work with CGI Group (CGI) to bring Azure into the Ft. Collins, CO, and Riverdale, MD datacenter locations. Doing so would allow a certified and accredited platform to exist inside of APHIS for its most sensitive data and pay for only what is consumed. The Microsoft Azure environment could be used when added capacity is needed for a short time or for experimentation for proof of concepts that programs may need to explore.

APHIS has adopted Microsoft Azure as its official cloud platform to enable APHIS programs to take advantage of its ability to expand capacity rapidly for existing Information Systems that could be ported.

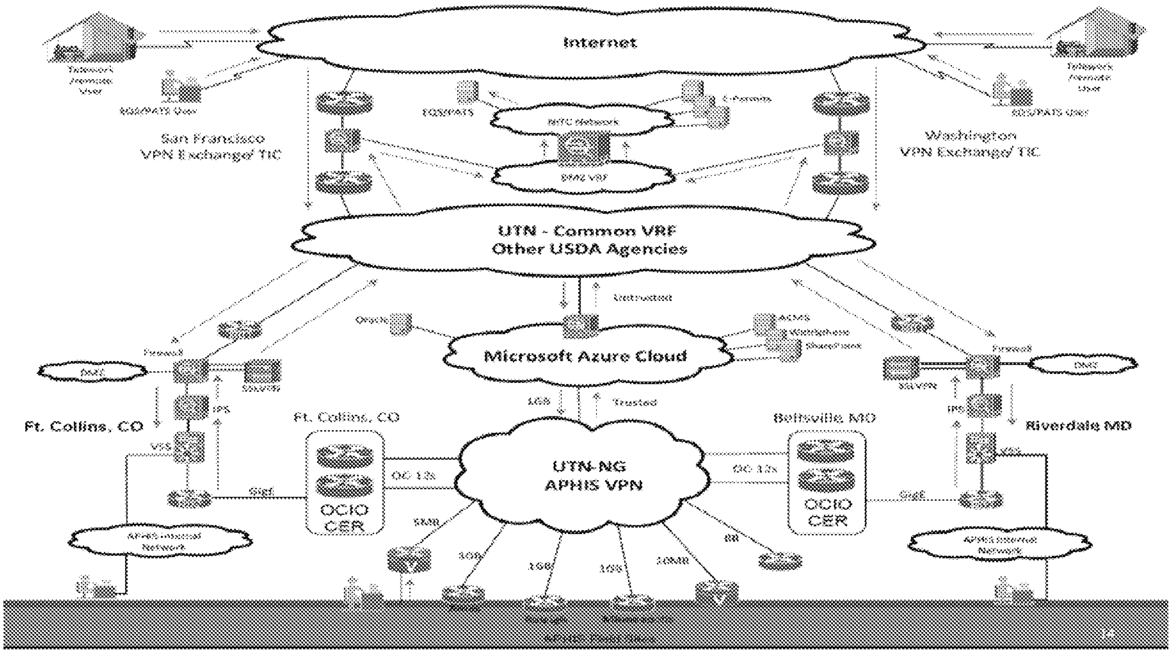
Microsoft Azure offers the ability for custom applications to be grown using native services instead of virtual images, allowing for more dynamic scaling than the confines offered by standard virtual machines.

> APHIS *has a sizable investment in Microsoft Azure and 1 Gbps connectivity between* APHIS *UTN and Azure already exists.*

APHIS has achieved Risk Management Framework (RMF) Phase I and has obtained an Authorization-To-Test (ATT) in the Microsoft Azure Cloud allowing initial workload testing in the cloud and refining migration plans. September 2017 is the planned pilot when APHIS will officially test the Azure platform.



APHIS has a converged architecture via Azure ExpressRoute via a 1 Gigabyte per second (Gbps) connection from APHIS UTN-NG VPN cloud.



Azure Stack

While being able to leverage the cloud side of Azure gives APHIS great flexibility, there are also tasks with additional security precautions and localized workloads that do not make sense to move into Azure but would benefit from Azures overall platform. Azure Stack benefits include:

- Build and deploy innovative applications using a consistent framework, processes, and tools across cloud and on-premises environments
- Extend Azure in the datacenter and enable on-premises Azure-consistent services
- Azure Stack brings the agility and innovation of cloud computing to on-premises environments, helping accelerate cloud adoption
- Scientists can use the same approach to build apps for Azure and Azure Stack, so applications can be deployed easily to either location based on regulations, the need to protect sensitive data, customization, and latency
- Application data can be kept where it belongs - in Microsoft datacenters, with a service provider, or in an APHIS datacenter
- Speed new application development building on components from the Azure Marketplace, including open source tools and technologies using the rich Azure ecosystem
- Take advantage of a continuous stream of new Azure technologies and services that support development efforts
- Implement Azure Stack enhancements with pre-validated updates applied to the entire integrated system

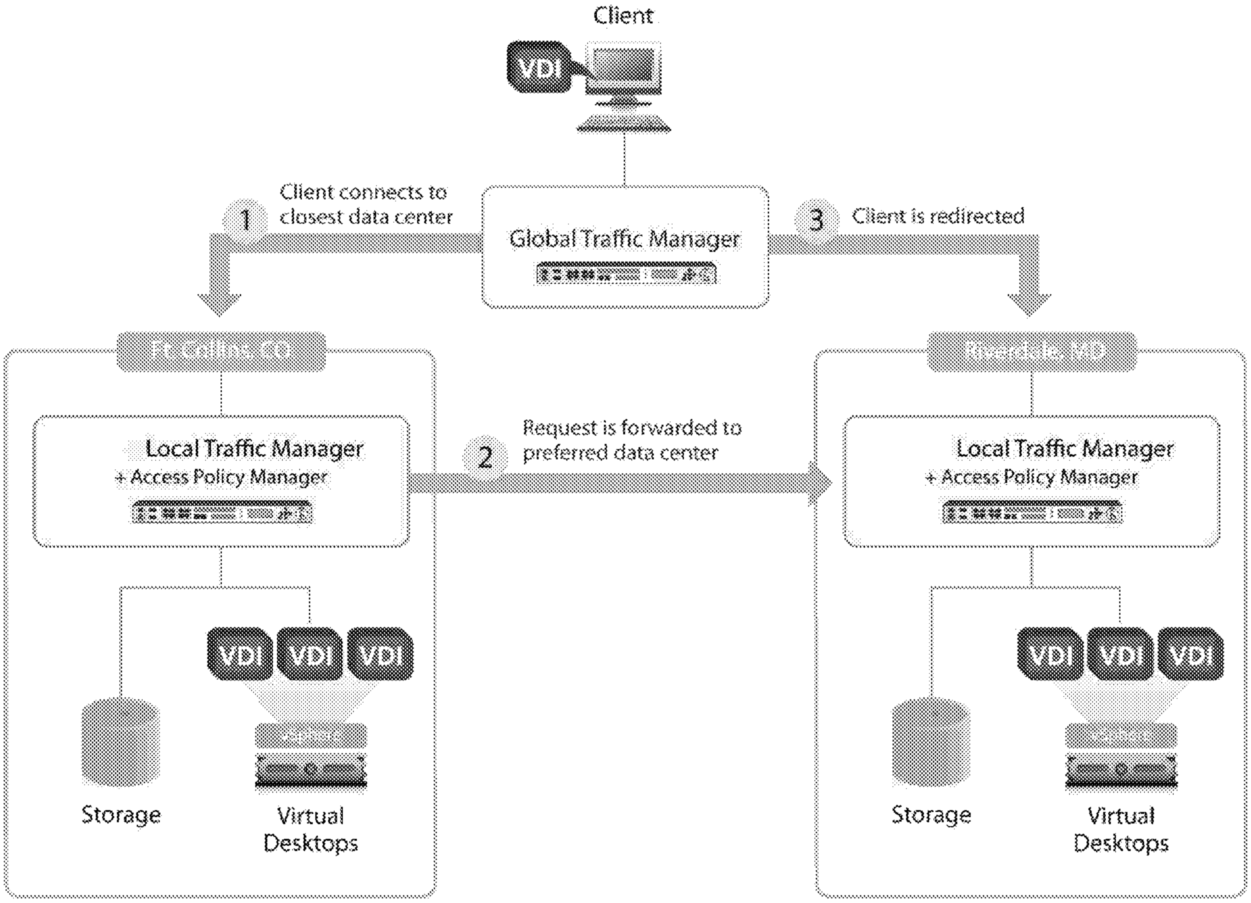
To avail these benefits, APHIS could purchase integrated systems from CGI under USDA DCOI and pay Microsoft for only the Azure services used.

Office Connectivity Challenges for Centralized and Cloud-Based Services

APHIS will have some challenges when it comes to office connectivity to centralized and cloud-based services. As discussed in SOF, network connectivity at some sites is so limited that large attachments in emails can bring those networks to a halt. While APHIS is working to modernize its footprint for the legacy T1 connections and bringing them up to 5 to 20 Mbps connections, there is still an issue with accessing large datasets for programs that are sitting outside of the Ames, IA, Ft. Collins, CO, and Riverdale, MD locations.

Virtual Desktop Infrastructure (VDI)

VDI would allow users in remote locations to use a desktop that would in effect, be in one of the two hub locations and access software to work on data as if it was located locally. VDI would allow remote users located in the office or on a VPN to have access to all of their applications without the need to have to transfer large amounts of data over the network.



APHIS already has VDI running on-premises in Ames, IA with over 150 current users running on VMWare Horizons 6.2. APHIS could replicate this footprint to Ft. Collins, CO and Riverdale, MD to improve continuity and support geographic affinity for sites that are closest to those Campuses.

APHIS sites that are data producers should point their instrumentation that generates data to Riverdale, MD or Ft. Collins, CO so that as the data is produced it is ready for access. In some instances, this may not be practical, and the continued practice of having to send data on portable media for data ingest will have to continue until such a time as the connectivity's capacity is increased. The end goal will be to store all the data for APHIS in its primary datacenters and afford users remote access to it.

VDI would allow APHIS to realize its vision to consolidate large datasets at Riverdale, MD, or Ft. Collins, CO datacenters.

An APHIS on-premises VDI would mitigate a large part of the current problem of data stored on individual systems that have no visibility by APHIS Enterprise IT. This approach allows APHIS to gain visibility into almost all of the datasets in the organization and take steps to understand the unique and individual needs of each program. A comprehensive understanding will facilitate solutions to unique program issues, and all of APHIS can take advantage of a centralized infrastructure.

Use or disclosure of data contained on this page is subject to the restriction on the title page of this proposal.

SCINet - Overview and Current State

While centralization of existing systems is underway at various sites, there is still a substantial need for scientific data and workloads to be handled in an environment that is uniquely equipped to deal with the problems of large computation needs, interchange data at high speed, and access specialized software. APHIS is already integrated with ARS at the Ames, IA and Ft. Collins, CO facilities and could leverage the ARS SCINet infrastructure to realize an immediate benefit in segregating scientific and enterprise computing environments.

The Science Network (SCINet) offers near-term SIT solutions including the option to extend building 580 in Beltsville, MD to the ARS campus via a 10 – 100GB connection.

SCINet is a national scientific network and research IT infrastructure that was established and is operated by ARS. SCINet currently spans six ARS locations across the US (Beltsville, MD; Ames, IA; Ft. Collins, CO; Albany, CA; Clay Center, NE; Stoneville, MS). SCINet is implemented on top of the Internet 2 (I2) Research and Engineering Network infrastructure and provides 10-100Gbps connectivity. SCINet provides ARS Scientists with a high-speed network for collaboration and data interchange. In addition to the high-speed network backbone, SCINet also features a high-performance computing (HPC) and storage infrastructure called Ceres. Ceres is housed in a data center in the shared ARS/APHIS National Centers for Animal Health (NCAH) facility in Ames, IA. The HPC infrastructure at NCAH is integrated into the SCINet network and provides ARS Scientists with the ability to store and analyze their scientific data, collaborate, and also conduct other types of scientific simulations. Ceres currently has 1,560 compute cores and is managed by the SLURM Workload Manager. High-performance storage is provided by a Seagate CS9000 Lustre storage appliance with a total of 1.3 PB of usable storage. Ceres also has two Data Transfer Nodes (DTNs) to facilitate data movement, collaboration, and data sharing that are configured specifically for high-speed data transfers into and out of SCINet.

Protected Data would have to be minimized and kept on Ceres purely for processing models such as multiple genetic sequences that would have no meaning until tied to the reports which could reside in an APHIS storage environment. A secure enclave that only APHIS has access to could be established and employ protections like Digital Loss Prevention and File Integrity Management to ensure tampering of these platforms would not be possible.

The VRSC provides ARS Scientists with research IT support and specialized subject matter support such as informatics and GIS. Support includes installation and optimization of scientific software, helping with user questions, and providing tutorials and regular user training.

In addition to direct access to the HPC environment via the command line, SCINet offers web-based services to leverage the HPC environment using browser-based tools geared toward novice users. An example is the web-based Galaxy bioinformatics platform that is currently available to ARS scientist within SCINet.

Ceres was designed with extensibility in mind to be able to adapt to the changing needs of ARS Scientists. Based on demand, additional compute capacity, storage, and hosted services can be added and provisioned as needed. ARS' SCINet effort has been underway for the last 36 months and has reached production with many ARS Scientists already using it and more being on-boarded regularly.

An important component of SCINet is the Virtual Research Support Core (VRSC) consisting of research IT staff and scientific domain experts.

SCINet Buildout

Although SCINet has reached production status, additional enhancements and integrations are planned over the next 1-2 years to bring SCINet to its intended state within ARS. While all six (6) external hub locations are connected into the SCINet network, Stoneville is the only site that has so far been fully integrated into SCINet - five out of the six locations still need to be fully integrated into the SCINet network. Work to fully integrate each of these locations includes integration of scientific data producers (instruments) or other scientific equipment (compute clusters, storage devices, etc.…) directly into SCINet depending on the scientific needs. At each location, a firewall will be deployed to enable direct access from the local ARSNET enterprise network to the local SCINet enclave. In addition, Local DTNs (LDTN) will be installed to facilitate data transfers within SCINet and other external locations. The integration strategy will be customized for each location and requires in-depth discussion with Scientists and IT staff at the site to determine requirements and needs.

At the Ceres HPC facility at NCAH, several enhancements to the HPC and networking infrastructure are planned that are intended to increase the performance, stability, and utility of the system for Scientists. A new 1Gbps commodity internet circuit is being deployed that will provide commodity access into and out of the SCINet environment. A key Ceres infrastructure need is for an additional, secondary storage tier beyond the existing high-performance Lustre storage system that is mainly intended for fast scratch storage. This second storage tier would host home directories and provide long-term data storage and archiving.

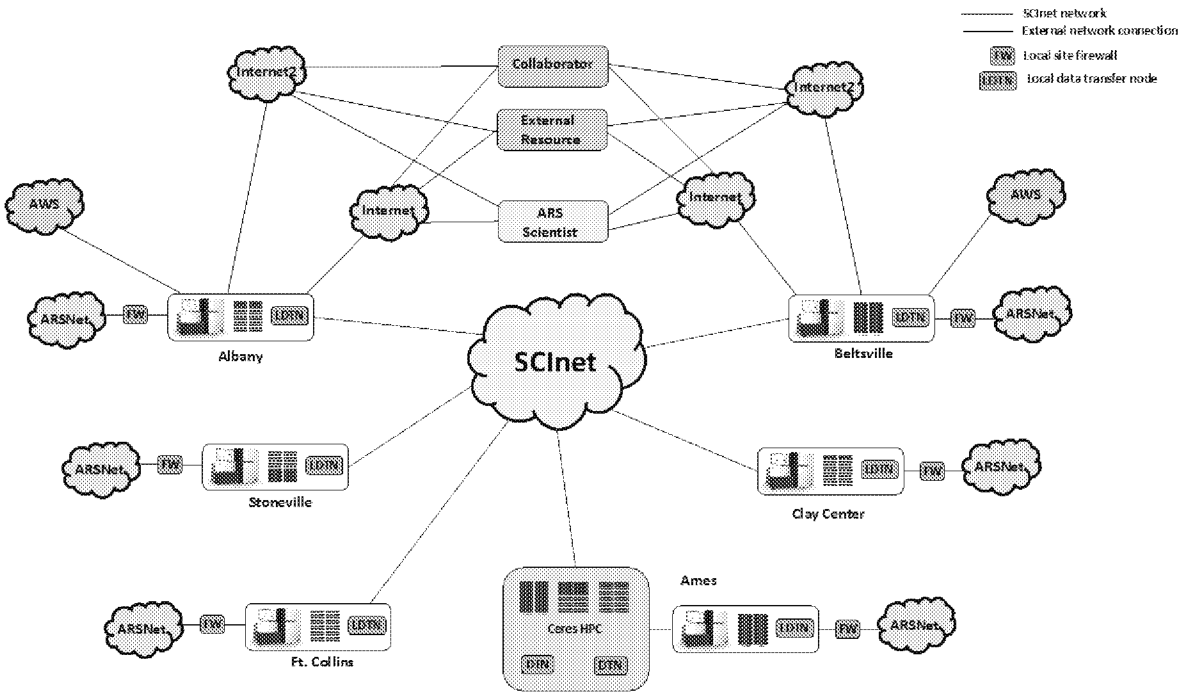
ARS staffs assisted by the DIFZ/BioTeam are working on the approval process for Authority to Operate (ATO) for the SCINet environment. This process is expected to take up to one year or longer and will be a key milestone for SCINet to become fully operational. ATO for SCINet may also be a key requirement for collaboration with other agencies such as APHIS.

While ARS SCINet has a public Cloud presence via Amazon Web Services (AWS), SCINet itself and the HPC infrastructure specifically are not yet integrated into AWS. Future AWS integration could consist of cloud bursting, cloud storage, or leveraging other AWS services such as databases or hosting services.

Finally, ARS is considering relocating the Ceres infrastructure from NCAH to a colocation facility on the ISU campus. DIFZ/BioTeam is working on providing ARS with an assessment strategy that will highlight the pros and cons of such a step. This strategy will include an assessment of the implications for ARS' ability to collaborate with other federal partners such as APHIS as part of the SCINet project.

Data Transfer within SCINet

SCINet's underlying network design is based on the Science DMZ concept originally developed by the Department of Energy's ESNet. Its network architecture is optimized to support efficient data transfer within SCINet itself as well as with external locations dependent on their available network bandwidth. The figure below shows a high-level logical view of the SCINet network topology including the six hub locations. A key component of SCINet's data transfer architecture is two dedicated DTNs that are part of the Ceres high-performance computing (HPC) infrastructure in Ames, IA. These DTNs have high-speed network interfaces and are tuned for high-speed data transfers within SCINet and also to outside locations. In addition, each of the six initial SCINet hub locations (Clay Center, Ft. Collins, Beltsville, Stoneville, Ames, and Albany) will feature a Local DTN (LDTN). The LDTNs will provide efficient data transfer within SCINet and in particular from the locations to the central DTNs in Ames. At the time of writing of this report, the integration work of the hub locations is ongoing and only one site, Stoneville, has been fully integrated into SCINet so far. The main DTNs in Ames enable efficient data transfer to and from external data sources such as collaborators or data providers. The depicted firewall between SCINet and the APHIS network is proposed but still in the approval process.



ARS is planning to deploy high-speed data transfer software, Globus, to leverage the DTNs as a key piece of infrastructure to facilitate data exchange across SCINet. In particular, the main DTNs would host a Globus endpoint. ARS Scientists and their collaborators would be able to leverage Globus for seamless and effective data transfer both within SCINet and with outside sources. In addition, other high-performance data transfer tools such as Aspera could be deployed, but there are currently no plans in place to do so. The advantage of Globus for data transfer versus command line driven Secure Copy (SCP) or File Transfer Protocol (FTP) is convenience and efficiency. Globus is optimized to maximize use of available bandwidth, e.g., by splitting the data transfer across multiple simultaneous network streams. In addition, users can queue their data transfer jobs, and Globus will perform the transfer without users needing to monitor or supervise the process. Data transfers could also be performed using more traditional software such as SCP but performance of these tools is typically significantly inferior to optimized protocols such as Globus or Aspera and not generally recommended.

To initiate a data transfer from an external hub location such as Clay Center to the Ceres HPC environment a user would select both the source dataset on local storage in Clay Center and the target data location on the Ceres file storage and then schedule the transfer. This process can be conducted via either the Globus web interface or Globus' CLI. Globus will ensure that data is transferred efficiently and notify the users on completion. ARS users who are not at one of the hub locations can still leverage Globus for data transfers between their location and SCINet. However, data transfer speeds will be limited by local connectivity into SCINet. ARS Scientists located at university locations that have SCINet connectivity can typically leverage it for effective data transfers. Otherwise, the commodity internet access into SCINet can be used for data transfers.

ARS Scientists who need to share data with external collaborators or have to pull data from external data producers into SCINet can also leverage Globus. The transfer speed is limited by the connectivity of the collaborator or data producer. Sources with SCINet connectivity will typically see the best transfer speeds; otherwise, the less performant commodity internet access into SCINet has to be used. This allows even ARS Scientists who are at a location with poor network connectivity to make efficient use of SCINet as long as there is a reasonable connectivity between SCINet and the target location itself; download raw data directly onto SCINet from the data producer (hopefully via a fast connection), process it on the Ceres HPC environment, and then pull back only the typically significantly smaller final result files to the local location over the slow connection.

SCINet Integration of Local Infrastructure

ARS Scientists work at ARS locations spread across every US state. These locations differ substantially in network connectivity as well as in the type of local IT infrastructure and scientific instrumentation. Locations that handle or produce significant amounts of scientific data (e.g., by running scientific instruments such as sequencers or mass spectrometers) or otherwise require efficient data exchange within SCINet to and from the Ceres HPC infrastructure or external collaborators, might be candidates for direct integration into SCINet by becoming a hub location in addition to the current six hubs. Direct integration of IT infrastructure into SCINet would typically provide the most efficient network path from that infrastructure to other SCINet locations including the Ceres HPC infrastructure in Ames. However, direct integration of a location into SCINet would require establishing a direct fiber path with SCINet's network backbone. Direct integration thus requires careful planning and may involve a substantial financial investment depending on the geographical location of the site to be integrated with respect to SCINet network hubs. In addition, additional networking gear such as firewalls, routers, or data transfer nodes may need to be purchased at directly integrated locations. For ARS sites that already have a SCINet presence (e.g., due to co-location at a major university), it may be more cost effective to forgo direct SCINet integration and instead leverage existing SCINet connectivity for access and data transfer into/out of SCINet.

ARS' chosen model for integration of external locations into SCINet is multi-pronged and dependent on the specific needs of individual locations, their existing infrastructure and connectivity, and cost.

Scientific Software within SCINet

Currently, more than 300 command line (CLI) tools are available on Ceres, and their number is growing steadily. Below is a small selection of tools available on Ceres:

- **Cufflinks** - Cufflinks assembles transcripts, estimates their abundances, and tests for differential expression and regulation in RNA-Seq samples
- **gatk** - the toolkit offers a wide variety of tools with a primary focus on variant discovery and genotyping
- **muscle** – Multiple Sequence alignment of protein and nucleotide sequences
- **raxml** – Randomized Axelerated Maximum Likelihood is a popular program for phylogenetic analysis of large datasets
- **tophat** - bioinformatics sequence analysis package tool for fast and high-throughput alignment of shotgun cDNA sequencing reads
- **velvet** - an algorithm package that has been designed to deal with de novo genome assembly
- **maker** - a portable and easily configurable genome annotation pipeline
- **samtools** – Sequence Alignment Map is a generic format for storing large nucleotide sequence alignments
- **blast** - Basic Local Alignment Search Tool is an algorithm for comparing primary biological sequence information, such as the amino-acid sequences of proteins or the nucleotides of DNA sequences

SCINet's Ceres HPC environment provides ARS Scientists with access to a substantial number of scientific tools, compilers, and scientific libraries.

If a particular tool is not available, SCINet users can contact the VRSC and request that the missing tool be installed. Software tools are made available on the Ceres HPC infrastructure via the widely used Environment Modules system. Ceres users can pick and choose exactly what type of software and version they need for their particular scientific use case. The software available on Ceres is diverse, ranging from serial scientific applications which run on a single CPU core to highly parallelized applications that can run across tens to hundreds of cores. The SLURM Workload Manager deployed on the Ceres HPC infrastructure is configured to allow most of these applications to run efficiently on the system. In practice, however, ongoing user training will be required to help ensure efficient use is made of the available HPC resources. SCINet also has a dedicated license server that can be used to serve licenses for scientific software tools that require them, e.g., via the flexlm license management system.

Finally, SCINet also has the ability to host scientific applications that use web-based or another graphical user interface to enable novice users to access the HPC infrastructure. Good examples are Science Gateways of the Galaxy web framework for bioinformatics which is available as part of SCINet for ARS users.

APHIS/ARS Integration Needs

ARS' SCINet is a purpose designed Research IT network and infrastructure aimed at supporting ARS Scientists, who are spread out across many different locations in each of the United States. From a Research IT and geographical distribution point of view there is substantial overlap between the needs of APHIS and ARS Scientists. Integration of APHIS into SCINet would benefit APHIS Scientists greatly. The closer collaboration of APHIS and ARS on SCINet would also enable cost sharing between the two agencies resulting in overall cost savings for both ARS and APHIS.

Short-term - use the facilities that already have SCINet

Log-term - integrate other facilities into SCINet or use VDI solutions to make them accessible immediately

A key requirement for closer collaboration between APHIS and ARS on SCINet would be a formalized agreement between the two agencies, which defines the nature of the collaboration as well as any requirements, Service Level Agreements (SLAs), or other guarantees that need to be in place. For example, APHIS may require additional and more stringent security controls than ARS around certain types of data that may impact the SCINet design and operating model. Once these requirements have been defined, understood, and agreed on they would guide and inform further evolution and design of the SCINet infrastructure. Depending on the specific nature of the agreements between ARS and APHIS, additional security controls that may be required by APHIS and the integration of APHIS into SCINet may impact the existing SCINet design and necessitate re-engineering of certain aspects of the system. The earlier these requirements can be understood and defined, the easier it would likely be to accommodate them in the still-evolving SCINet ecosystem.

From a technical point of view, the requirements for integrating APHIS Scientists or locations into SCINet would be similar to the ones currently faced by ARS. APHIS locations that need to be integrated directly into SCINet as hub locations would need fiber connectivity into the SCINet network backbone as well as investment into local networking equipment (routers, firewalls, etc.). APHIS locations that are collocated with existing ARS hub locations could likely be integrated more easily and inexpensively compared to stand-alone APHIS locations.

Data Storage

APHIS has a variety of storage platforms located on the Ft. Collins, CO, Ames, IA, Riverdale, MD, and Raleigh, NC campuses:

Location	SAN	Capacity TB	% Allocated
Ft. Collins, CO	Pillar	271	92%
	IBM	21	48%
	Tintri	13	2%
	FalconStor	65	100%
Ames, IA	Pillar	330	90%
	Dell	36	83%
	Dell	24	92%
	FalconStor	306	74%
Riverdale, MD	Pillar	236	79%
	Tintri	13	2%
	FalconStor	61	100%

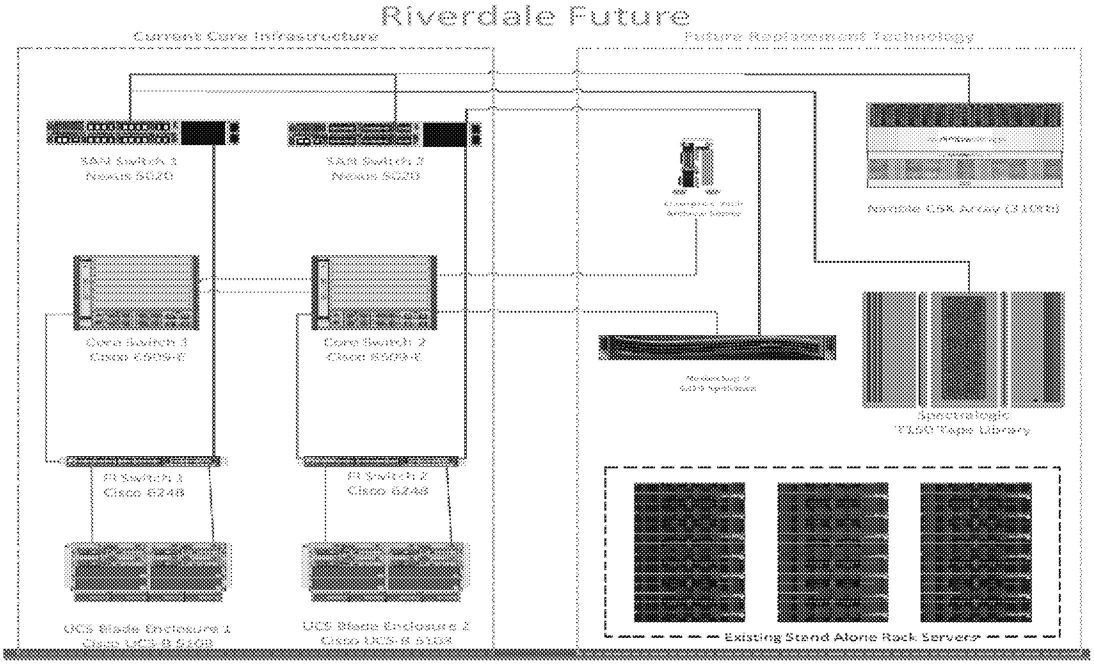
Location	SAN	Capacity TB	% Allocated
Raleigh, NC	Pillar	20	50%
		1,396	**68%**

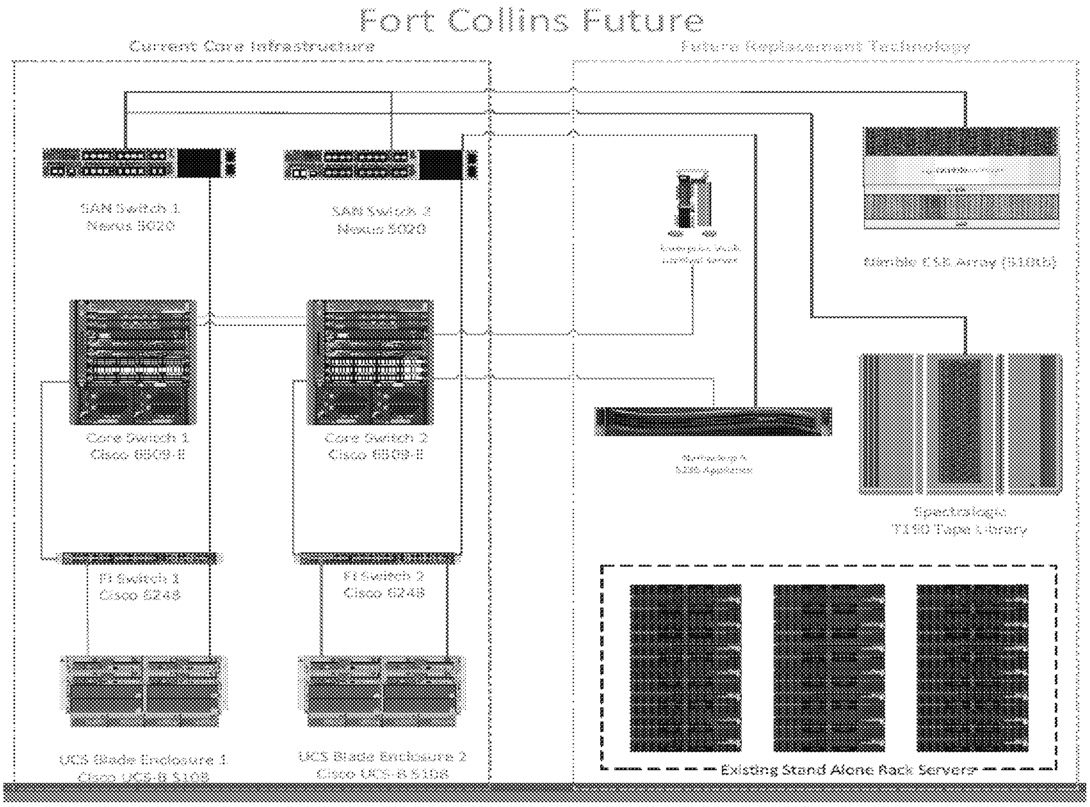
Overall APHIS has 1,396 TBs of storage with 68% overall utilization. The FalconStor and Tintri storage which was used by the VDI solutions are planned to be retired, and approximately 30 TB of vSAN storage is being used on the organic Dell servers in the Ames, IA and Ft. Collins, CO site.

APHIS has identified Nimble Storage as the next solution for the enterprise. The Scientists and IT staff at the Ames, IA facility want to utilize the CS9000 Luster Parallel Filing System solution from Seagate to take advantage of its unique parallel growth scalability capabilities. Ames, IA Scientific datasets are around 76TB and growing at a rate of 235% annually. While the ARS SCINet Luster file system has its advantages, there may be issues with some datasets that require Confidential Information Protection and Statistical Efficiency Act (CIPSEA) protections and must reside in a properly accredited environment. Extra protections may be needed, or a small enclave carved out in SCINet to allow the data to reside there and remain compliant with APHIS and USDA security practices.

The Oracle Pillar Axioms SAN is at end-of-life and is to be replaced by Nimble Storage for the enterprise platform due to its significant speed and throughput advantages over the current Oracle Solution.

The figures that follow show the planned APHIS storage platform post migration to Nimble Storage at the Riverdale, MD and Ft. Collins, CO facilities.

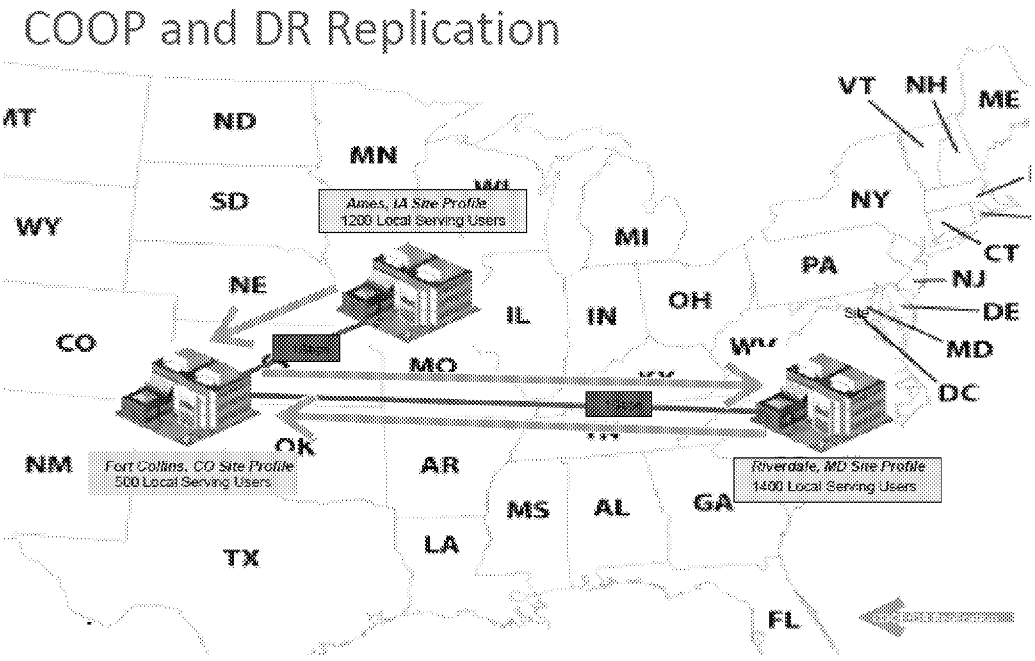




Storage Replication Topology

The Oracle SAN was originally set up to replicate between Ames, IA and Ft. Collins, CO. However, issues of bandwidth and the sheer volume of data have led to backups having to be diverted to tape and shipped between the locations to ensure data continuity. Plans to re-initiate the inter-site replications will occur after APHIS updates its storage solution to Nimble.

Once APHIS consolidates SAN technology and establishes inter-site replication data centralization can begin.



Centralizing Storage from Field Sites

Servers that are targeted for elimination at the individual sites around APHIS will account for approximately 213 TB of data that may need to be relocated to the centralized storage platform. If the storage that is removed from individual sites is relocated at either East or West, then access to the data and applications can be granted via the proposed VDI solution discussed earlier in this section.

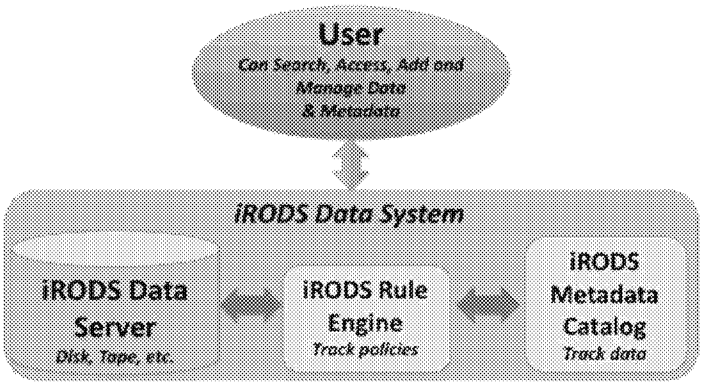
Metadata Indexing

While growing storage capacity does meet the needs for growth of APHIS' datasets and enterprise data, it does not address all of the facets of data retention. Retention policies alone are not enough to help govern what data is stale, what could be removed, and what should be kept in the long run.

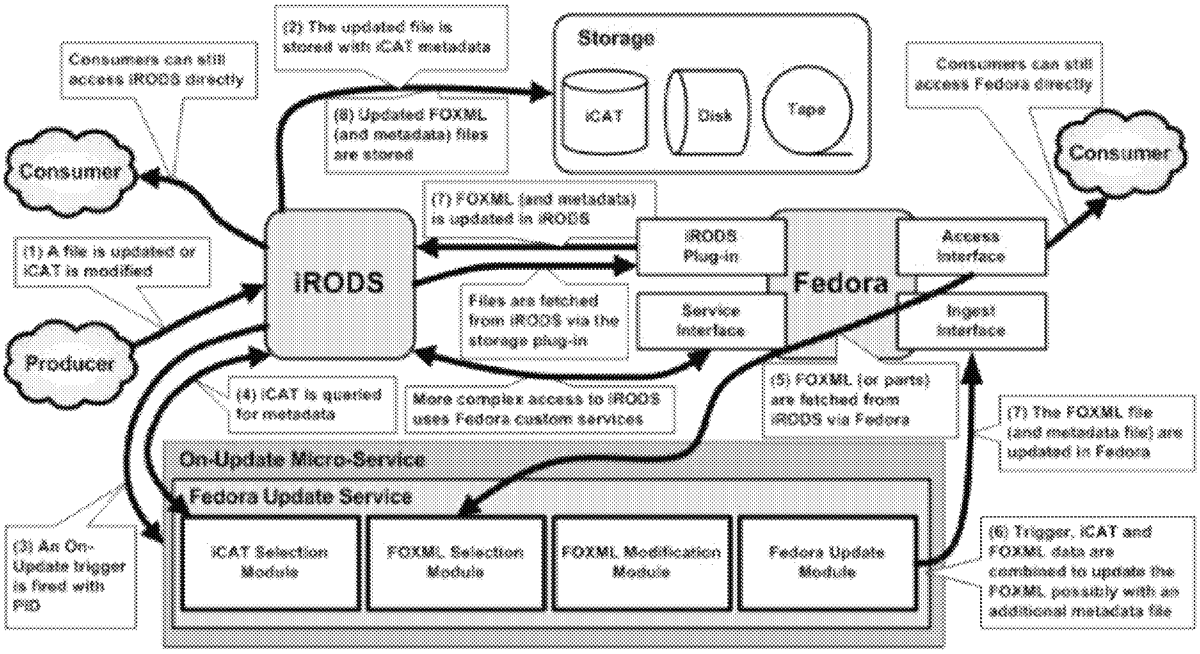
All too often, organizations with vast scientific computing and instrumentation are constantly adding data to their storage platforms. Couple this with constant churn in personnel or projects and it is clear how datasets grow invalid over time. Determining datasets that are relevant and the ones that are stale, is a key component to being able to expunge aging data that is no longer valid.

> *APHIS programs should seek to integrate as much data as possible into a metadata indexing system like iRODS, which provides database content indexing.*

Integrated Rule-Oriented Data System (iRODS) allows Scientists to assemble collections from data that resides anywhere in the enterprise or scientific computing environments. Creating indexed collections and defining what they contain enables other users and groups to know and understand the purpose and nature of those datasets. This understanding is particularly helpful for APHIS when engaging in cross-agency or organization based collaboration.



This platform will control policies attached to the data through procedures that allow access control and data integrity. The iRODS system provides journals of all actions to support auditing systems to capture where the data has gone and who has accessed or possibly changed the datasets.


Consumers can still access iRODS directly
(2) The updated file is stored with iCAT metadata
Storage
iCAT
Disk
Tape
Consumers can still access Fedora directly
Consumer
(8) Updated FOXML (and metadata) files are stored
Consumer
(7) FOXML (and metadata) is updated in iRODS
(1) A file is updated or iCAT is modified
iRODS
iRODS Plug-in
Fedora
Access Interface
Files are fetched from iRODS via the storage plug-in
Service Interface
Ingest Interface
Producer
(4) iCAT is queried for metadata
More complex access to iRODS uses Fedora custom services
(5) FOXML (or parts) are fetched from iRODS via Fedora
(7) The FOXML file (and metadata file) are updated in Fedora
On-Update Micro-Service
Fedora Update Service
(3) An On-Update trigger is fired with PID
iCAT Selection Module
FOXML Selection Module
FOXML Modification Module
Fedora Update Module
(6) Trigger, iCAT and FOXML data are combined to update the FOXML possibly with an additional metadata file

The Geospatial Information Systems (GIS)

APHIS has a substantial investment in Esri's ArcGIS and ArcMap based software that is installed locally on workstations in most APHIS programs. Although APHIS Enterprise IT is pushing for the use of Esri's cloud-based app, most programs that have tried this complain about usability.

Some of the basic APHIS standard build workstations being used are inadequate for the computational demands being put on them. For example, Wildlife Services has a shapefile that has millions of data points so when they try to use Esri's portal, the application crashes. When used on a workstation that contains 32GB of RAM, it can take minutes to re-render after the user pans the position on the map.

We recommend that APHIS conduct a study to determine the types of workloads that are suitable to migrate into Esri's cloud-based application versus keeping them on the desktop application. Furthermore; the study should include whether or not that current hardware is sufficient for existing workloads. APHIS should seek a server class system with multiple Graphics Processing Units (GPUs) to assist in rendering large GIS datasets.

Looking into the future, workloads can be virtualized in the cloud. Understanding how intensive those workloads will be determines the type of cloud-based system that will be required to support them.

2.3 Long-Term Strategies

This section concentrates on long-term opportunities that exist for Scientific and Hybrid Cloud Computing. The following table cross-references four of the identified issues and opportunities to evaluated and proposed long-term strategies.

Long-Term Strategies	Hybrid-Cloud	Metadata Indexing	Storage
High-Performance Computing	✓	✓	
Big data and metadata storage and management	✓	✓	✓
High-speed communication for data sharing and collaboration	✓		✓
Program Level Requirements	✓	✓	✓

2.3.1 Program-Sponsored Solutions

The program concerns documented in SOF have database and scientific processes unique to each program. Some of the programs have similar needs such as data portability between sites or external collaborators due to limitations on network speed. Computational power has been a concern for a majority of the programs whether it is for running models, Geospatial Information, or Sequencing.

Storage

Storage is a chief program concern. CNSS found that individual Scientists and programs keep some data on laptops and workstations and often backup that data to portable storage media, which range from simple memory sticks to large 1 TB external hard drives. The sum of the entire catalog of scientific data is kept in this manner. Loss of these systems could cause catastrophic impact to the scientific programs.

Some APHIS programs already implement Western Digital (WD) Network Attached Storage (NAS) devices to centralize local data backups. The WD My Cloud PR4100 (https://www.wdc.com/products/network-attached-storage/my-cloud-pr4100.html#WDBNFA0400KBK-NESN) has a range of storage options up to 40 TB that are relatively separate from Enterprise IT solutions.

Each site that stores data locally would benefit from centralizing the backups on a local storage appliance that could then be backed up to enterprise storage. This approach would be especially recommended for sites that do not contain access to Enterprise Storage locally.

Once APHIS has realized its long-term plans to update its centralized IT storage and has optimal connectivity to each remote site using the WD My Cloud PR4100, the storage devices could be shipped back to APHIS to be absorbed into the centralized storage. Once the storage is absorbed/consolidated, the WD My Cloud PR4100 could be kept in inventory to migrate data or for field operations. While not directly in-line with the enterprise storage strategy, this is a cheap stop-gap measure that will enable consolidation of disparate storage devices to ensure critical failure of these devices won't critically impact programs.

While this solution may not be ideal for the long-term APHIS storage strategy, it affords the sites that do not have good connectivity to establish a centralized storage platform. This could be used as a stop gap measure to consolidate various storage devices on-site to mitigate failure of those devices. In the long-term, the connectivity will be upgraded or a VDI solution will be in place to allow the data on-site to be accessed remotely.

The WD My Cloud PR4100 solution is low cost enough to make this short-term solution disposable.

Veterinary Services (VS)

VS is well-resourced in Ames, IA and Ft. Collins, CO, with robust technical support from knowledgeable IT staff dedicated to enhancing the scientific and Enterprise IT environment. Location does afford the program access to SCINet and is the easiest of all programs to migrate due to the colocation.

VS has already run network switches from SCINet into their labs and Café in Ames, IA. However, they have not deployed workstations that would connect directly to SCINet for the purposes of kiosk based data transfer from VS to SCINet. This is intended to aide Scientists for migration of data at high speeds into SCINet with minimal difficulty.

VS suffers from the same storages issues as all APHIS programs. Scientists and program personnel are storing data on portable storage devices. Connectivity to remote sites like Plum Island, NY is also inadequate so portable storage media also serve to transport data sets between sites, partners, and programs when network transfer is prohibitive. VS IT staff are working to modernize the storage platform located in Ames, IA and Ft. Collins, CO. Once this platform is modernized, there will be less dependency on portable media as all data will reside in SCINet or the local Enterprise IT storage platform.

VS has a need to conduct metadata indexing for the purposes of tracking its data and moving away from the standard distributed file system it is using today. The current method of data storage makes finding data sets difficult and often leads to duplication of data and potentially the lingering of data sets long after their usefulness. Portability of data between partners and programs is also problematic. Metadata Indexing with a platform such as iRODS will allow VS to not only know where data resides, but also provide better visibility of the program's data as information on data's complete lifecycle would be present

VS has two Red Hat Enterprise Linux servers in Ames, IA and one in Plum Island, NY. While these systems are large capacity, they are not joined as an HPC platform. This can be established via the same OpenHPC (https://openhpc.community/) solution that is described in the Wildlife Services (WS) section. The small cluster of workstations in Ft. Collins, CO would also benefit from OpenHPC and maximize the workloads executed on that end of life lower-end hardware.

VS has a heavy reliance on Open Source software and also has a need to collaborate internally on its development. While VS is using public facing free GIT repositories, they have a need for internal management of their software and models while collaborating with partners and fellow Scientists. GitLab (https://about.gitlab.com/) is a platform that could be implemented internally or in the cloud for greater collaboration.

Biotechnology Regulatory Service (BRS)

BRS conducts field assessments and maintains the data in the ePermits system as described in SOF. One of the primary issues and concerns for BRS is the outdated regulatory forms downloaded onto mobile devices used at the inspection site to capture data. On return from an inspection, data must be entered manually into the ePermits system as there is no automated import mechanism.

BRS is working on the old technology with the ePermits system and would benefit from an application customized to work in an inspection capacity that could be run on mobile technologies. Previously such applications have required a substantial investment and constant maintenance. However, FileMaker Pro (http://www.filemaker.com/) allows data to be presented as an application across any platform with no coding or programming skills. FileMaker Pro would allow field personnel to create new, flexible applications around the existing datasets and manage data formats from a simple and easy to use interface.

The House of Representatives and many other Government agencies are using FileMaker Pro to capture data from the field and extend access to internal data from a simple to use interface with no programming.

BRS could implement FileMaker Pro and still interact with the already established ePermits database because FileMaker Pro replaces the web-based front end of ePermits while leaving the overall infrastructure intact so that Business Intelligence solutions such as Cognos could continue to be used for data visualization and reporting purposes.

Wildlife Services (WS)

WS has a host of issues identified in SOF. Addressing these issues with program-centric solutions that are independent of APHIS Enterprise solutions has yielded some important results.

IT Collaboration & Outreach

WS needs to coordinate with VS as they have done a great job with their IT outreach program where Scientists and IT professionals meet to discuss how technology could be used to improve and enable scientific programs. WS would benefit greatly from this type of program. While WS is mostly regulatory, this model allows the IT staff who report to the Enterprise IT department to act as advocates for WS' needs with a deep understanding of the mechanics of the program and resulting needs.

Search Engine Optimization

A major issue is that the WS public-facing search engine is effectively broken as it does not return relevant search results; some of the responses from the search engine are decidedly random. WS contracted a search engine which yielded relevant search results and minimized superfluous data to an outside vendor for a time, but the program was terminated.

The key mechanic that led to the success of this program was metadata injection. While the iRODS system would allow WS to take advantage of these features, WS believes that their needs are currently too small for such a solution unless it is driven from the Enterprise IT department.

However, WS is working with the U.S. Forestry Service (USFS) as they have been USDA leaders in making their data and archives publicly accessible.

WS should continue to work with USFS or possibly extend their archives into the USFS system to benefit from their enhanced methods.

Air-Gapped Environment

WS has an air-gapped environment on their Foothills campus that contains many reclaimed workstations used as local computational resources. The facility room where these computers are located suffers from heating and cooling issues, causing them to be able to use only a subset of the computers. These issues can be solved easily by dividing up the computers in multiple rooms or investing in the facility to accommodate the added electrical load.

WS uses these workstations as single computational units, which limits the collective power of the machines. Options do exist to transform the individual units into a High Performance Computing (HPC) environment.

Although cluster implementation is relatively simple, minor adjustments to scripts that run the models may have to occur to implement this HPC environment, and there may be an IT skills gap for the local WS IT personnel. Training Scientists to work in an HPC environment may be required. However, the results would pay large dividends in a localized HPC platform and get scientist's code ready for implementation into the larger SCINet platform.

OpenHPC is a project that allows any Linux distribution to be able to create an HPC cluster. This approach allows more efficient and quicker computations to occur when running scientific languages like R.

Archive Integrity Management

WS has a large 5TB and growing archive that requires a high level of manual maintenance to ensure file system integrity is maintained. Pushing all WS data onto a read-only file system would not guarantee that corruption would not occur as the information would still have to be manually hashed by the open source utility, droid.

File Integrity Monitoring (FIM) is an internal control or process that can monitor every file transaction to validate the integrity of operating systems and application software files using a verification method comparing the current file state and a known, trusted baseline. FIM automates integrity monitoring in the file system thereby ensuring archives remain in a trustworthy state and reducing the overall hours it takes to manage an archive holistically. WS archives that will benefit from FIM include:

- NAIS Traceability Archives
- NWRC Archives
- Wildlife Scientific Data Archives
- Wildlife Tissue Archives
- Stakeholder Meetings Archives

FIM Solutions

While these solutions have a security-centric flavor, Integrity is one of the most important tenants of security and maintaining a viable archive. Ensuring that files aren't changed due to manipulation or data corruption is one of the most important aspects to ensure the data contained in archives remain in a trustworthy state.

SolarWinds File Integrity Monitoring (http://www.solarwinds.com/topics/file-integrity-monitoring) with Log and Event Manager offers key features for security, compliance, and troubleshooting:

- Fast and easy compliance reporting
- Real-time event correlation
- Real-time remediation
- Advanced search and forensic analysis
- File integrity monitoring
- USB device monitoring

WS would enhance productivity greatly by investing in one of several (FIM) products, which scale from enterprise software to open source.

Tripwire File Integrity Manager (FIM) (https://www.tripwire.com/products/tripwire-file-integrity-manager/) provides the ability to integrate File Integrity Manager with many APHIS security controls: security configuration management (SCM), log management, and Security Information and Event Management (SIEM). Tripwire FIM adds components that tag and manage the data from these controls intuitively and in ways that protect data. For example, the Event Integration Framework (EIF) adds valuable change data from File Integrity Manager to Tripwire Log Center or almost any other SIEM. With EIF and other foundational Tripwire security controls, you can easily and effectively manage the security of your IT infrastructure.

OSSEC (https://ossec.github.io/) is an Open Source Host-based Intrusion Detection System. It performs log analysis, integrity checking, Windows registry monitoring, rootkit detection, real-time alerting and active response. It runs on most operating systems, including Linux, OpenBSD, FreeBSD, Mac OS X, Solaris, and Windows.

Offline Scientific Equipment

WS has some scientific measuring equipment that is network and cloud-enabled. Due to security concerns, WS does not connect these appliances to the network for the software updates. Instead, the appliances are maintained in an offline state, and the manufacturer sends a technician to update them manually. This process is both time-consuming and a cost burden for WS.

One way to support remote updates without compromising network security would be to allow only certain traffic to beacon to the device when an update is initiated. This approach would take some careful study to allow the systems to be accredited so the level of risk to the rest of the environment would be minimized by concise security controls implemented during this process. Once updates occur, devices could be returned to an operational offline state.

WS and APHIS Enterprise IT should work to get instruments connected to the network so they can be updated remotely.

The most optimal solution for WS would be to work towards the network devices to be online full-time so that they can communicate with the manufacturer and obtain updates. This would ensure the devices are at the latest firmware and enhance overall network security.

WS may benefit from using the same technique as VS. VS connects their devices through a windows server or workstation, using it as a router to access online updates which are then able to be downloaded onto the scientific instruments. This keeps the scientific instruments off the network directly which mitigates any direct communication with them to exploit vulnerabilities. This allows the scientific instruments to remain online and cloud connected without the security risk to the network.

Plant Protection and Quarantine (PPQ)

PPQ is one of the largest programs in APHIS and is spread across all states. CNSS conducted interviews and documented the major program components in SOF and has found that the following recommendations would benefit the PPQ program as a whole, and address unique areas of emphasis by site or function.

CPHST & PGQP Beltsville, MD Lab

Outside of VS, building 580 has the most scientific instruments that support their mission, making them the largest data producer in PPQ at this time. While some of the instruments are not connected to the network, they do produce data and are integral in sequencing raw data that is then analyzed and turned into finished products.

CPHST & PGQP are using mostly reclaimed workstations that are not connected to the APHIS network to run their models and execute custom code to conduct the analysis. Since these systems are not maintained by APHIS Enterprise IT, they present a single point of failure that could result in catastrophic data loss. The labs are also utilizing a wide array of portable media to store critical data. Again, this could lead to catastrophic data loss if those devices become corrupt, lost, or are accidentally destroyed.

CPHST & PGQP are good candidates for the WD My Cloud PR4100 and addition of a Server configured like VS Ames, IA, and Plum Island sites.

A WD, My Cloud PR4100 type of NAS, onsite would address the storage needs and allow the labs to grow their data onsite while awaiting Enterprise solutions that will eventually move the data off of the site into the hub sites (Ft. Collins, CO, and Riverdale, MD). This solution would allow them to grow up to 40TB in size relatively inexpensively and afford them cloud backup to ensure data integrity and availability in the event of a disaster.

Since building 580 is such a large data producer and substantial analysis is done onsite, PPQ is acquiring a server with similar specifications as the servers in VS at the Ames, IA and Plum Island sites. The server would allow the lab to move off end-of-life workstations and onto a central platform where tools and models could be run.

Open source tools are up and running in the VS environment, they have not been incorporated at the Beltsville, PPQ lab because of security concerns. CPHST & PGQP should continue to work with Enterprise IT and security to get these needed tools so they could take advantage of the onsite server platform for centralized computations.

Otis Labs

Otis Labs is performing numerous complex data modeling tasks where the ArcGIS shapefiles are in excess of 3GB. Otis Labs is unable to utilize the cloud version of Esri's ArcGIS as loading the entire dataset crashes the portal. The only way to get the cloud version to work is to drop out a majority of the data points or make the area rendered smaller and incrementally sew results together to create a complete product.

Utilizing the workstations at Otis Lab, the file can be loaded fully, but simply panning the map causes the file to take several minutes to load making a simple task that should be relatively quick last hours. The workstations that are supporting this process are standard enterprise Dell Laptops with 16GB of ram. The issue is that ArcGIS is a highly computational intense system for rendering imagery and requires a substantial system with multiple 3D graphics capabilities.

CNSS recommends that PPQ and Otis Labs look into getting Precision 5720 All-in-One Professional workstations with the following specifications:

Intel® Xeon™ Processor E3-1275 v6 (Quad Core HT 3.8Ghz, 4.2GHz Turbo, 8MB)
2 x AMD Radeon Pro WX 7100 w/8GB GDDR5 Graphics Cards
64GB (4x16GB) 2133MHz DDR4 ECC Memory
1TB M.2 PCIe SSD Class 50 Solid State Hard Drive

All PPQ locations where ArcGIS computational power is needed would benefit from the recommended workstation platform.

CPHST Mission Lab, PPQ

CPHST is on its way to being a large data producer like its Beltsville, MD counterpart. The issue at this time is that the specimens are sent to external labs and then downloaded in massive datasets in excess of 100GB in size. To accomplish this, Scientists have to go to the nearby University of Texas Rio Grande Valley Campus and utilize the I2 connection to download these datasets in a timely fashion.

While this is a perfectly accessible solution given the proximity of the university and availability of the tools, it is a less than convenient scenario to accomplish the downloads. This may be less of an issue when NextGen Sequencers come online, and they start sequencing locally. They will still need to share, and archive data and the current 20 Mbps lines are inadequate for today's purposes.

Some alternatives would be:

- Get the connectivity from Internet 2 from the nearby university extended to the lab
    - This would also afford this site direct access to SCINet with coordination with ARS
    - Sequencing could then be delivered directly into SCINet as would the follow-on analytics
- Continue to outsource the sequencing to external labs or even another PPQ building 580
    - This would allow the Mission Lab to utilize VDI solutions outlined in the Near-Term solution to remotely administer their data on centralized tools as if it were localized on site without the need to upgrade the 20 Mbps connection

VDI is a more out of the box solution to leverage the current processing power and high-speed network connectivity at the hub sites.

These alternatives would work well since SCINet is the desired environment for APHIS to be conducting scientific processing.

PGA Message Set

The PGA Message Set has been an issue for PPQ for some time now. It has been mentioned in Congressional Reports as an issue for PPQ and remains a compliance issue. The message set contains data that is derived from the Automated Commercial Environment (ACE) and is delivered via email every 5 minutes. The current dataset has expanded to about 50 GB in size and contains over 15 million rows of data.

Right now, this data resides in a flat file and is visualized with a business intelligence tool Power BI (https://powerbi.microsoft.com/en-us/) on local workstations and in Azure Cloud. PPQ has a desire to make this dataset more static by importing it into a SQL Server Data Warehouse either locally (MS SQL Express) or through Azure Native SQL Services. The data could then be connected to other data sources like treatment systems (head, cold chemical) and the anti-smuggling system database. The fact that the dataset may be inter-connected with larger datasets in the future and is growing at more than 50GB per year makes this dataset a candidate to be classified as "Big Data". Its growth and interconnection cites a need to ensure that the dataset is properly implemented to tier for growth and more frequent use by a wider audience.

Tableau (https://www.tableau.com/) is another platform that could provide visualization for the PGA Dataset and is already used in the greater APHIS environment. Other open-source products like RapidMiner (https://rapidminer.com/) allow similar types of access to the data visualization.

Power BI, Tableau and RapidMiner products all provide some level of business intelligence to allow the mining of datasets. While APHIS has a considerable investment in Tableau, smaller datasets may be better served by solutions like Power BI and RapidMiner. These solutions have a lower cost and should be fully explored to see if they meet the unique data visualization needs of the program.

Spatial Analysis Framework for Advanced Risk Information Systems (SAFARIS)

SAFARIS currently reside at the North Carolina State University (NCSU) campus on an APHIS provided server. While this platform is not currently considered Big Data, it does house a large quantity of PDF files from statistical analysis that does currently take up several terabytes of space. CNSS recommends that this server be integrated into the APHIS Enterprise network at Riverdale, MD where the proper storage can be allocated in the Enterprise Storage. Virtual Desktops could be employed at Raleigh to allow users of SAFARIS to access and manipulate its data from their site or remotely. This would allow for the APHIS vision of consolidation to meet the DCOI mandate.

2.3.2 Consolidated APHIS/ARS Cloud

SCINet is a focal point within our near-term solutions section. SCINet already exists at Ft. Collins, CO, and Ames, IA where ARS and APHIS collocate. To get SCINet in a production state used by ARS and APHIS Scientists alike, important long-term issues that need to be solved are:

- Storage of APHIS data in SCINet
- Access to SCINet in APHIS facilities via a firewall
- SLA or Memorandum of Understanding (MOU) between Agencies

Storage

The bulk of the scientific data would need to reside in SCINet for processing. It is costly and time-consuming to move large datasets between APHIS and SCINet because of APHIS privileged data security policy. Either privileged data needs to be redacted from the datasets, or an enclave within SCINet needs to be established so that APHIS data can reside in SCINet without violating security policy. Given that it is difficult and costly to move the datasets continuously between environments makes this the most desirable solution.

CNSS recommends that APHIS move to adopt SCINet as its primary scientific computing platform and the establishment of a secure enclave to keep data in that environment makes the most sense.

Access

Access to SCINet is difficult at this time because it is an air-gapped environment not currently connected to the APHIS network. There are access points on the Ames, IA and Ft. Collins, CO campuses that could be used to give Scientists access to move their datasets into SCINet via community workstations.

One such workstation does exist in Ames, IA, but has not been configured or placed in the Café at the facility. Switches have been extended into many labs and the Café which could provide more standalone access points in the facility where Scientists could bring their datasets to ingest into SCINet. This same model could be extended to all facilities such as PPQ's building 580 in Beltsville, MD to allow data ingestion into SCINet.

These sites could serve as data ingest sites, and all facilities that do not have access to SCINet would have to send their datasets on portable storage media for data ingestion.

SLAs

Establishing SLAs is one of the most important tasks for APHIS Enterprise IT. The SLAs will determine what APHIS is entitled to use and how many resources at a given time. Clauses could be inserted into the SLAs dealing with national emergencies or the need to surge resources in SCINet that could suspend or kick jobs running on the Ceres HPC platform.

Going forward, APHIS and ARS need to work together as the SCINet platform is extended or enhanced to gain full befit of its resources.

2.3.3 APHIS-only Public & Hybrid Cloud

CNSS has combined both the Public & Hybrid cloud since most of the services and architecture are seamless to either solution either by onsite virtualization or public cloud instances.

CNSS has explored additional cloud options aside from Microsoft Azure, the APHIS Enterprise IT preferred platform. While this platform may be the right choice to offload enterprise workloads and simplify enterprise IT infrastructure, it remains to be a proven platform conducive to scientific computing. For further exploration, CNSS has researched services and migration options for large datasets into the Amazon Web Services (AWS) platform and the partially implemented APHIS Microsoft Azure Platform. Each cloud platform is explained in-depth below. The discussion also evaluates how each can be leveraged in order for APHIS to access their building blocks to truly unlock the power of the cloud.

While each platform has a series of unique services that it offers, the real power of the platform is in leveraging its unique building blocks to allow business units and Scientists to break down very complex architectures or processes into a series of simple inputs that work together instead of porting individual systems with all the software built in. This involves a shift in thinking of how the existing platforms, datasets, software, and scientific models are put together and how they must be completely re-built for the cloud. The industry term for this is "Infrastructure as Code," and it sets your processes free versus keeping them locked into a virtual machine with only finite resources.

> *Building applications, processes and even scientific models for the cloud will yield a nearly unlimited scale in processing and overall cost savings as you pay for only the resources applications use. In the long run, rebuilding apps allows access to what would potentially be millions of dollars in capital expenses for mere pennies on the dollar.*

ARS already has connectivity established to AWS through SCINet via Internet 2. APHIS may be able to leverage this connectivity at its Ames, IA and Ft. Collins, CO facilities to move its scientific datasets into AWS for processing much in the same way it would move data in and out of SCINet. This would alleviate the issue that AWS does not yet have a direct or dedicated connection established in AWS via the USDA Unified Telecommunications Network (UTN). Even in the event of establishing a connection through the USDA UTN, in all likeliness it would be limited to 1 Gbps of speed much like the Azure established connections. This is not optimal for large datasets that would need to be transferred in and out of the platform.

While not an overnight process, if done correctly, Scientists and Enterprise Users alike will find that breaking free of traditional infrastructure and being able to scale processes in a nearly unlimited way makes the adoption worth the effort in the long term.

AWS

AWS provides several valuable services which would meet and scale to the needs of APHIS Scientists and employees. Amazon Machine Image (AMI) is one of these services and can provide the information required to launch an instance, which is a virtual server, in the cloud. AMIs are available for most tools utilized by APHIS and those that are not can have a custom AMI created.

AWS is the recognized industry leader in public cloud hosting and has the security, flexibility, and scale to meet the requirements of the APHIS migration effort. Use of AWS does not restrict APHIS from using other public clouds such as Microsoft Azure, IBM Soft layer, or several other cloud hosting providers may be utilized in a manner similar to the techniques described below. For this reason, we reference options for Microsoft Azure and AWS in this document.

> *Due to its large geographic footprint, wide range of services, and low cost, CNSS proposes the AWS cloud as a scientific infrastructure provider either in conjunction with ARS or on its own for the migration and hosting of Big Data.*

Federal Risk and Authorization Management Program (FedRAMP) and AWS GovCloud

To provide end-to-end security and end-to-end privacy, AWS builds services in accordance with security best practices, provides the appropriate security features in those services, and documents how to use those features. The AWS cloud infrastructure (including non GovCloud US Regions) has been designed and managed in alignment with regulations, standards, and best-practices including Service Organization Controls (SOC) 1/Statement on Standards for Attestation Engagements (SSAE) 16/International Standard on Assurance Engagements (ISAE) 3402 (formerly Statement on Auditing Standards [SAS] No. 70)

- SOC 2
- SOC 3
- Payment Card Industry Data Security Standard (PCI DSS)
- International Organization for Standardization (ISO) 27001
- ISO 9001
- ISO 27001
- Department of Defense Risk Management Framework (DoD RMF) Cloud Security Model (CSM)
- Federal Information Security Management Act (FISMA)
- Federal Information Processing Standard (FIPS) 140-2
- Family Educational Rights and Privacy Act (FERPA)

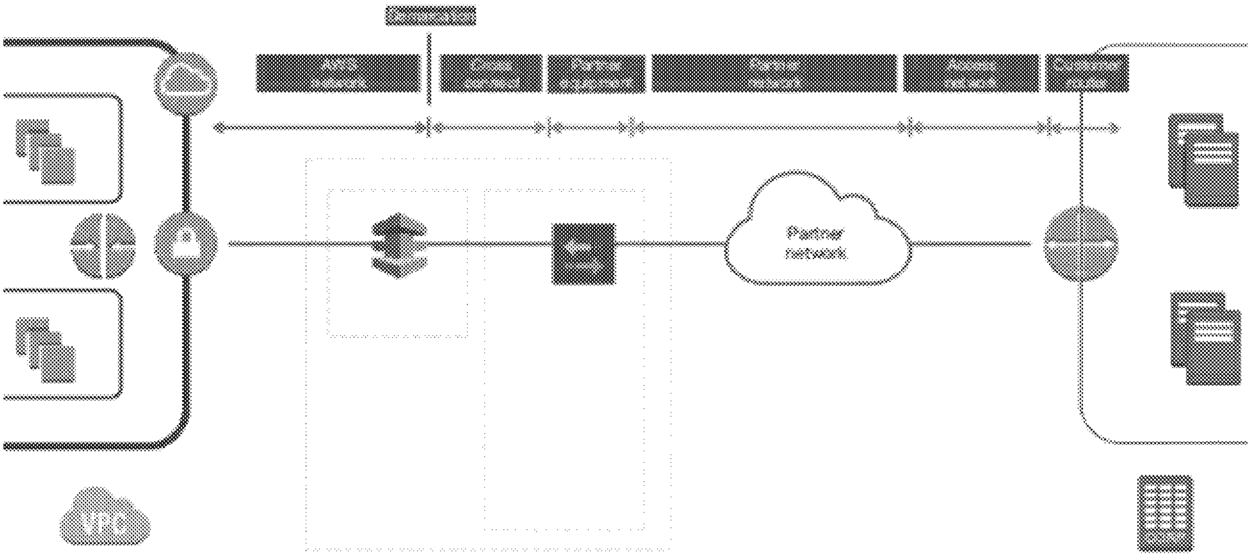
AWS will provide the SOC1 report to Customers under NDA. The AWS security center provides up to date information on AWS audits by independent third-party auditors.

AWS US Regions adhere to FedRAMP compliance. All US regions are SOC complaint, and SOC reports can be provided on request. Should APHIS require ITAR compliance, GovCloud should be used. Otherwise, we recommend building APHIS resources in the US West Oregon Region due to its lower cost, greater accessibility, and greater breadth of services. For the purposes of high-level cost estimates, we have used AWS GovCloud Region pricing in most cases. However, our recommendation is to use US West Oregon Region instead of GovCloud whenever possible, and likely for all of the migration. Using the commercially available cloud service such as the AWS US-West 2 (Oregon) region will save on average:

- 17% off Server Cost
- 47% off Long Term Storage (Glacier)
- 23% off Object Storage (S3)
- 17% off EBS Storage (Block Storage)
- 23% off Glacier Transfer Costs
- 41% off S3 and other Networking Costs

AWS Direct Connect

CNSS recommends APHIS further evaluate using AWS Direct Connect, which lets you establish a dedicated, private network connection between your network and one of the AWS Direct Connect locations. Likely issues could be SLAs and bandwidth.



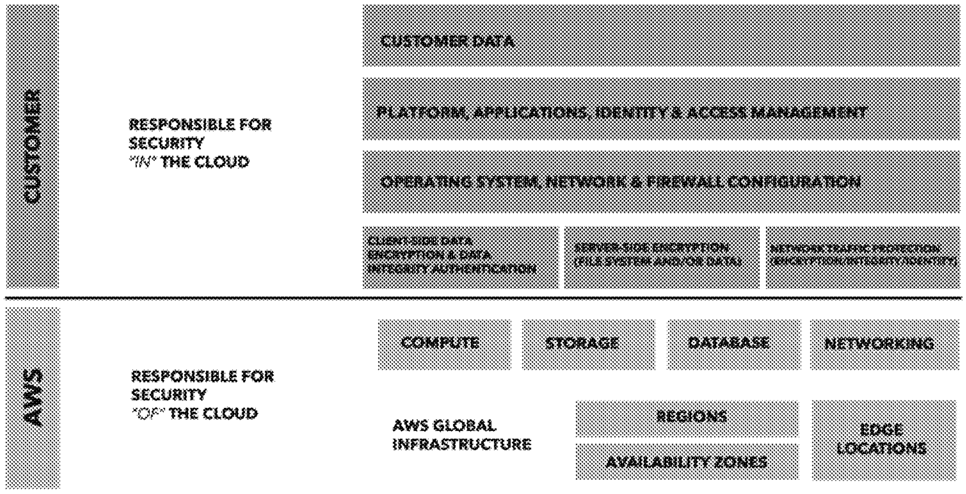
AWS Direct is ideal for hybrid cloud solutions and migration efforts that require increased bandwidth and a consistent, dedicated connection to the AWS resources. Direct connect enables the AWS environment to appear as an extension of the on-premises data center. It allows for secure and private traffic between on-premises and cloud servers and allows for secure VPN access into the cloud environment for server administration. For redundancy, a second Direct Connect can be provisioned, or an Internet-based VPN tunnel can be established.

Built for Security

The security posture of the environment will be similar to that of the on-premises datacenter. Security can be controlled via an on-premises firewall such as the already utilized Juniper firewalls or may be maintained by using the service provider's security features and virtual firewalls. The preceding figure represents the shared security responsibility model that outlines APHIS responsibilities on the AWS platform and platform provided security:

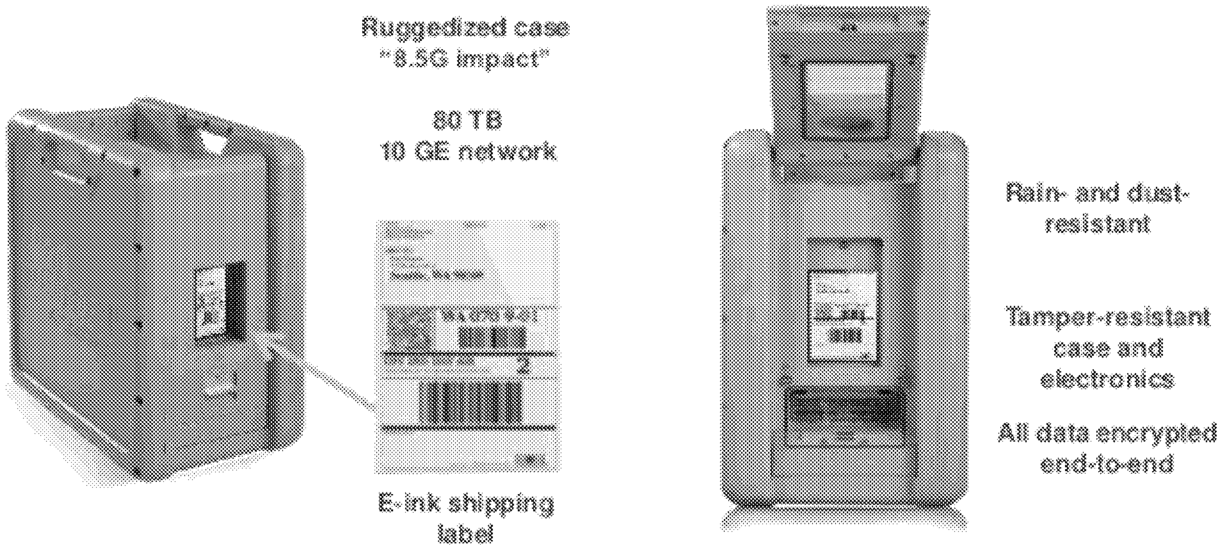
Numerous Virtual Private Clouds (VPCs) may be created with strong security between the VPCs. Each VPC may contain one or more subnets and expand between numerous datacenters (availability zones) for maximum scalability.

Routing within a VPC is enabled by default but may be modified and controlled to the firewall between subnets. Routing between VPCs is disabled by default and must explicitly be enabled. As an added security measure, no VPC may be a transit pathway between two other VPCs.


CUSTOMER
RESPONSIBLE FOR SECURITY "IN" THE CLOUD
CUSTOMER DATA
PLATFORM, APPLICATIONS, IDENTITY & ACCESS MANAGEMENT
OPERATING SYSTEM, NETWORK & FIREWALL CONFIGURATION
CLIENT-SIDE DATA ENCRYPTION & DATA INTEGRITY AUTHENTICATION
SERVER-SIDE ENCRYPTION (FILE SYSTEM AND/OR DATA)
NETWORK TRAFFIC PROTECTION (ENCRYPTION/INTEGRITY/IDENTITY)
AWS
RESPONSIBLE FOR SECURITY "OF" THE CLOUD
COMPUTE
STORAGE
DATABASE
NETWORKING
AWS GLOBAL INFRASTRUCTURE
REGIONS
AVAILABILITY ZONES
EDGE LOCATIONS

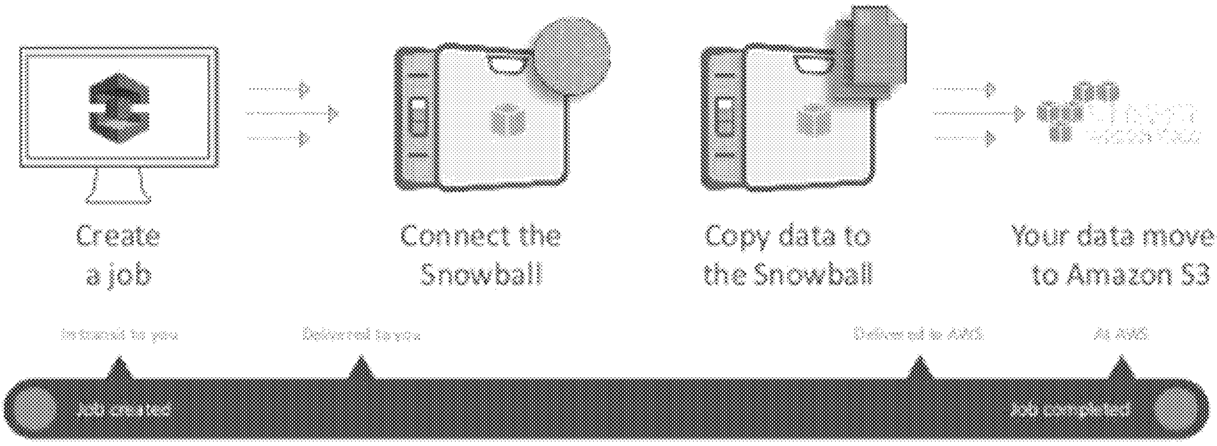
AWS Snowball – Peta-Scale Data Transport

For backups and migrating very large datasets to the cloud, Snowball is a highly valued Amazon service with the ability to transport near 1 PB of data into the cloud in 7 days without the needs of very expensive 10/100 GB lines between locations. Snowball is a storage device which is used for the physical transfer of data from one place to AWS locations. The device is essentially a large external encrypted data repository which can vary in size between 50 to 80 TB.



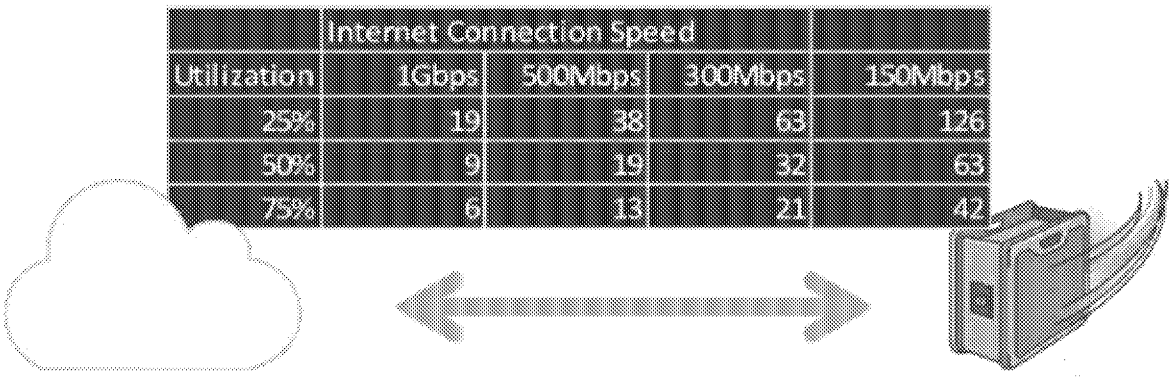
The Data stored on Snowball is stored in AWS S3 and Backup services. The use of Snowball would allow APHIS to transfer large amounts of data that needs to be accessible through AWS cloud services and backed through physical transfers. AWS Snowball provides a solution to several critical challenges including data storage, retention of data, transfer speeds, equipment, and accessibility.

Snowball Process:



- APHIS can order Snowball devices and sizes.
- AWS delivers Snowball devices to a physical address.
- APHIS connects the device to a network and downloads desired data on the Snowball.
- On completing upload, APHIS sends the Snowball back to AWS
- On receiving the Snowball, AWS begins the upload process into S3 cloud storage and Backup (if desired).

With AWS Snowball, transfers of 50 TB via a 10 Gbps connection can be completed to the cloud in less than one week, including shipping. For comparison, the figure below shows how long it would take to transfer 50 TB of data over traditional network connections on a scale of days:



	Internet Connection Speed			
Utilization	1Gbps	500Mbps	300Mbps	150Mbps
25%	19	38	63	126
50%	9	19	32	63
75%	6	13	21	42

Direct connect, another AWS option, would be a peer-to-peer connection between APHIS and AWS. The direct connection provides several benefits including:

- Reduced bandwidth
- Consistent network performance
- Elasticity
- Simplicity

Snowball uses for APHIS

APHIS programs have varying requirements for data retention length, as well as what types of data need to be retained. Snowball provides the ability to store all required data in the cloud. Since the Snowball is a physical device shipped between APHIS and Amazon, there are no issues with expensive network speeds.

AWS Import / Export Service

AWS Import/Export is a service that accelerates transferring data into and out of AWS using physical storage appliances, bypassing the Internet and is an alternative to snowball for getting smaller datasets into AWS. AWS Import/Export Disk was originally the only service offered by AWS for data transfer by mail. Disk transfers data directly onto and off of storage devices you own using the Amazon high-speed internal network. APHIS would achieve large data ingest by using the Amazon Import/Export Service eliminating the need to mail large mobile storage devices.

Infrastructure as Code or the DevOps Approach

The application of the principles we have discussed does not have to be limited to the individual resource level. Since AWS assets are programmable, you can apply techniques, practices, and tools from software development to make your whole infrastructure reusable, maintainable, extensible, and testable.

Automation in AWS

In a traditional IT infrastructure, you have to react manually to a variety of events. When deploying on AWS, there is abundant opportunity for automation so that you improve both your system's stability and the efficiency of your organization.

- **AWS Elastic Beanstalk** is the fastest and simplest way to get an application up and running on AWS. Developers can simply upload their application code, and the service automatically handles all the details, such as resource provisioning, load balancing, auto-scaling, and monitoring
- **Amazon EC2 Auto recovery**: You can create an Amazon CloudWatch alarm that monitors an Amazon EC2 instance and automatically recovers it if it becomes impaired. A recovered instance is identical to the original instance, including the instance ID, private IP addresses, Elastic IP addresses, and all instance metadata. However, this feature is available only for applicable instance configurations. Please refer to the Amazon EC2 documentation for an up-to-date description of those preconditions. In addition, during instance recovery, the instance is migrated through an instance reboot, and any data that is in-memory is lost.
- **Auto Scaling**: With Auto Scaling, you can maintain application availability and scale your Amazon EC2 capacity up or down automatically according to conditions you define. You can use Auto Scaling to help ensure that you are running your desired number of healthy Amazon EC2 instances across multiple Availability Zones. Auto Scaling can also automatically increase the number of Amazon EC2 instances during demand spikes to maintain performance and decrease capacity during less busy periods to optimize costs
- **Amazon CloudWatch Alarms**: You can create a CloudWatch alarm that sends an Amazon Simple Notification Service (Amazon SNS) message when a particular metric goes beyond a specified threshold for a specified number of periods. Those Amazon SNS messages can automatically kick off the execution of a subscribed AWS Lambda function, enqueue a notification message to an Amazon SQS queue, or perform a POST request to an HTTP/S endpoint
- **Amazon CloudWatch Events**: The CloudWatch service delivers a near real-time stream of system events that describe changes in AWS resources. Using simple rules that you can set up in a couple of minutes, you can easily route each type of event to one or more targets: AWS Lambda functions, Amazon Kinesis streams, Amazon SNS topics, etc.
- **AWS OpsWorks Lifecycle events**: AWS OpsWorks supports continuous configuration through lifecycle events that automatically update your instances' configuration to adapt to environmental changes. These events can be used to trigger Chef Recipes on each instance to perform specific configuration tasks. For example, when a new instance is successfully added to a Database server layer, the configure event could trigger a Chef recipe that updates the Application server layer configuration to point to the new database instance

- **AWS Lambda Scheduled events**: These events allow you to create a Lambda function and direct AWS Lambda to execute it on a regular schedule

Service, Not Servers

Developing, managing, and operating applications, especially at scale, requires a wide variety of underlying technology components. With traditional IT infrastructure, companies have to build and operate all those components. AWS offers a broad set of computing, storage, database, analytics, application, and deployment services that help organizations move faster with lower IT costs.

Manager Services

AWS offers a set of services, which provide building blocks that Scientists can consume to power applications. These managed services include databases, machine learning, analytics, queuing, search, email, notifications, and more. For example, with the Amazon Simple Queue Service (Amazon SQS) you can offload the administrative burden of operating and scaling a highly available messaging cluster while paying a low price for only what you use. Not only that, Amazon SQS is inherently scalable. The same applies to Amazon S3 where you can store as much data as you want and access it when needed without having to think about capacity, hard disk configurations, replication, etc. In addition, Amazon S3 can also serve static assets of a web or mobile app, providing a highly available hosting solution that can scale automatically to meet traffic demands.

> *It is possible to build both event-driven and synchronous services for mobile, web, analytics, and the Internet of Things (IoT) without managing any server infrastructure.*

There are many other examples of available services such as Amazon CloudFront for content delivery, Elastic Load Balancing (ELB) for load balancing, Amazon DynamoDB for NoSQL databases, Amazon CloudSearch for search workloads, Amazon Elastic Transcoder for video encoding, and Amazon Simple Email Service (Amazon SES) for sending and receiving emails, and more.

Serverless Architectures

Another approach that can reduce the operational complexity of running applications is that of the serverless architectures. These architectures can reduce costs because you are not paying for underutilized servers, nor are you provisioning redundant infrastructure to implement high availability.

When it comes to mobile apps, there is one more way to reduce the surface of a server-based infrastructure. You can utilize Amazon Cognito so that you don't have to manage a back-end solution to handle user authentication, network state, storage, and sync. Amazon Cognito generates unique identifiers for your users.

Those can be referenced in your access policies to enable or restrict access to other AWS resources on a per-user basis. Amazon Cognito provides temporary AWS credentials to your users, allowing the mobile application running on the device to interact directly with AWS Identity and Access Management (IAM) protected AWS services. For example, using IAM, you could restrict access to a folder within an Amazon S3 bucket to a particular end user.

For IoT applications, traditionally organizations have had to provision, operate, scale, and maintain their servers as device gateways to handle the communication between connected devices and their services. AWS IoT provides a fully managed device gateway that scales automatically with your usage, without any operational overhead for you.

AWS Workflows

For complex workflows, AWS offers Simple Workflow Service (SWF). Amazon SWF helps Scientists build, run, and scale background jobs that have parallel or sequential steps. Amazon SWF as a fully-managed state tracker and task coordinator in the Cloud. SWF will allow APHIS to track the state of processing, and recover or retry if a task fails. Using SWF with Amazon's Simple Queuing Service (SQS) will allow APHIS to build powerful highly resilient and robust multistep workflows.

Amazon SWF promotes a separation between the control flow of your background job's stepwise logic and the actual units of work that contains your unique business logic. Separation allows APHIS to manage, maintain, and scale "state machinery" of your application from the core business logic that differentiates it. As APHIS business requirements change, APHIS can easily change application logic without having to worry about the underlying state machinery, task dispatch, and flow control.

Amazon SWF runs within Amazon's high-availability data centers, so the state tracking and task processing engine is available whenever applications need them. Amazon SWF redundantly stores the tasks, reliably dispatches them to application components, tracks their progress, and keeps their latest state.

Amazon SWF replaces the complexity of custom-coded workflow solutions and process automation software with a fully managed cloud workflow web service. SWF eliminates the need for developers to manage the infrastructure plumbing of process automation so they can focus their energy on the unique functionality of their application.

Amazon SWF seamlessly scales with APHIS application's usage. No manual administration of the workflow service is required as you add more cloud workflows to your application or increase the complexity of your workflows.

Amazon SWF lets APHIS write application components and coordination logic in any programming language and run them in the cloud or on-premises. You pay a small charge for workflow executions in SWF. However, the first 1,000 are free. For more information, please see: https://aws.amazon.com/swf/

AWS Waiver

While conducting meetings with the AWS Federal Representative, CNSS discovered that a waiver is in place for research and academic institutions. The waiver can provide a maximum discount of 15% of the total monthly spending on AWS services. APHIS programs could receive this waiver either by being a research institution or by associating with academic institutions. APHIS would need to apply for the waiver through AWS.

APHIS may qualify for an AWS Egress waiver and be able to use parts of the platform for free.

AWS requirements for waiver of Egress charges:

- Work in academic or research institutions.
- Run any research workloads or academic workloads. However, a few data-egress-as-a-service type applications are not allowed under this program, such as massively online open courseware (MOOC), media streaming services, and commercial, non-academic web hosting (web hosting that is part of the normal workings of a university is allowed, like departmental websites or enterprise workloads).
- Route at least 80% of their Data Egress out of the AWS Cloud through an approved National Research and Education (NREN) network, such as Internet2, ESnet, GÉANT, Janet, SingAREN, SINET, AARNet, and CANARIE. Most research institutions use these government-funded, dedicated networks to connect to AWS while realizing higher network performance, better bandwidth, and stability.
- Use institutional e-mail addresses for AWS accounts.
- Work in an approved AWS Region.

Azure Solutions

As noted previously, APHIS has adopted Microsoft Azure as its official cloud platform in order to help meet DCOI compliance and enabling resources for APHIS programs to take advantage of its ability to rapidly expand capacity for ported existing Information Systems.

APHIS is not targeting Azure for porting virtual machines, but to take advantage of its unique platform abilities that will allow on-demand growth. Like AWS, Microsoft Azure offers the ability for custom applications to be grown using native services allowing for more dynamic scaling than what the confines of standard virtual machines offer.

The status of the APHIS Azure initiative and the architecture are discussed in more detail in the preceding Microsoft Azure and Azure Stack sections of this report.

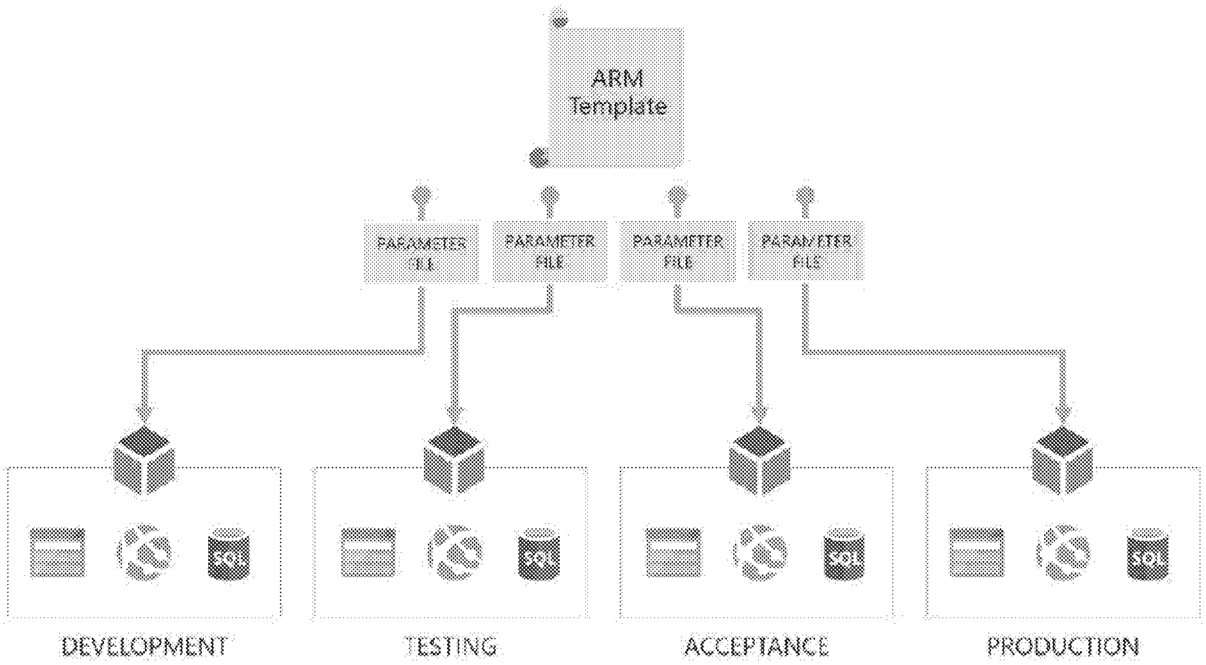
Infrastructure as Code or the DevOps Approach

APHIS needs to manage and operate their existing applications in a modern way, while also taking advantage of the cloud model whenever possible. This transformation shifts APHIS from a traditional model to a cloud model.

In the traditional model, applications are configured manually with scripts, user interfaces, management utilities or likely a combination of all these tools. The result of this process is an application that is released to a specific environment. During the lifecycle of the application different tooling manages different aspects of the application. When changes are made, they are performed in the management tooling, locally on the resource or in related resources that may be shared with other applications. Deploying two identical instances of an application with the same scripts on the same date will most likely not have identical configurations over their complete lifecycle, even if that is the desired state.

To overcome these challenges, the concept of Infrastructure as Code was introduced. It allows you to define the desired state of your application in a template. The template is then used to deploy the application. The template provides the ability to repeat a deployment exactly, but it can also ensure that a deployed application stays consistent with the desired state defined in the template over time. If you want to make a change to the application, you will make that change in the template. The template can then be used to apply the desired state to the existing application instance over its complete lifecycle.

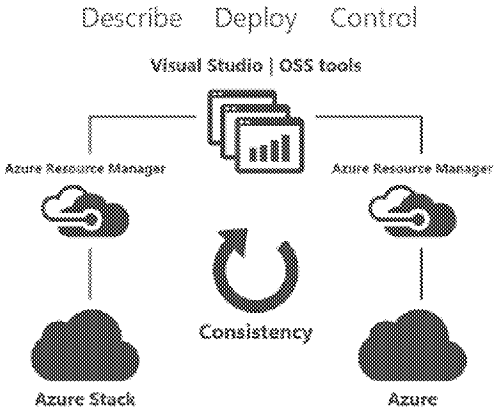
Templates can be parameterized; creating a reusable artifact that is used to deploy the same application to different environments, accepting the relevant parameters for each environment.



Benefit from Consistent Application Development

APHIS is able to realize value faster when the programs can build and deploy applications the same way, whether the apps run on Azure or Azure Stack. The Azure Resource Manager enables the same application model, self-service portal, and Application Program Interface (APIs). With the Azure Stack Development Kit, you can also support a robust dev/test environment on a single server.

APHIS can implement a common DevOps approach across a hybrid cloud environment and use common processes and tools across Azure and Azure Stack:



- Unified deployment experience with Visual Studio.
- Continuous integration/ continuous deployment (CI/CD) pipeline with open source tools (e.g., Jenkins) and Visual Studio Team Services (VSTS).
- Automation using Chef and Azure PowerShell DSC extensions.
- APHIS is able to benefit from the power of the 'One Azure' ecosystem. Speed up new cloud application development by using a range of open-source and community-driven software from the Azure Marketplace. Choose from multiple Linux distributions, Docker-integrated Containers (Linux and Windows Server), BlockChain, and Mesos.
- Use Cloud Foundry across Azure and Azure Stack to build and run cloud applications that are easily portable across hybrid cloud environments.
- Work with your choice of open source application platforms, languages, and frameworks including Java, Python, Node.js, and PHP.

Run Azure Services On-premises

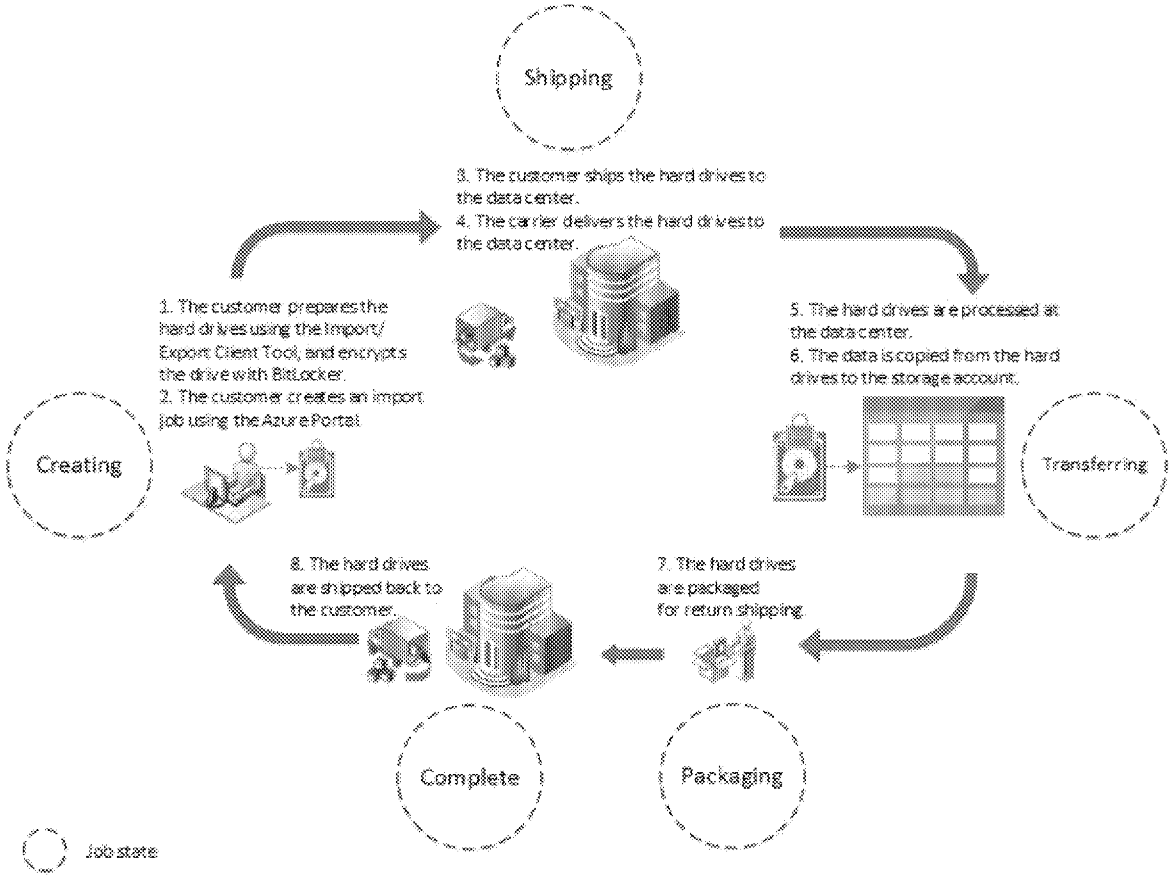
With Azure Stack, APHIS can run IaaS and PaaS services with the same administrative experiences and tools your team uses with Azure.

- Consistent Azure application platform services enable hybrid deployment choice and portability for cloud applications. Run PaaS (Azure App Service), serverless computing (Azure Functions), distributed micro services architectures (Azure Service Fabric), and container management (Azure Container Service) in on-premises environments.
- Consistent Azure IaaS services go beyond traditional virtualization. Virtual Machine Scale Sets, for example, enable rapid deployments with true auto-scaling for modern workloads such as containerized applications. You can also use Azure services to integrate Azure Stack into your datacenter.
- Management: Use Azure management and security services as your unified management solution, including for protection and disaster recovery of applications and workloads running on Azure Stack. (You can also use System Center Operations Manager Management Pack).
- Identity. Use your Azure Active Directory (AAD) subscription to administer Azure Stack identities, including secure multi-tenant access, which enables users across multiple AAD tenants to access Azure Stack resources.
- Azure integrated delivery experience
- With APHIS Azure Stack integrated systems, delivered by CGI under USDA DCOI, APHIS programs can focus on optimizing applications and services. Get up and running quickly with purpose-built Azure Stack integrated systems that you can choose from Dell EMC, HPE, Lenovo, with Cisco Systems. These systems come fully ready to run and offer consistent, end-to-end customer support.
- Pre-validated software updates delivered on a predictable schedule enable you to benefit from continuous innovation available from Azure. Add new services and additional Azure Marketplace applications.

Azure Import/Export Data Transfer Service

The Azure, Import/Export service, allows you to transfer large amounts of data securely to Azure storage by shipping hard disk drives to an Azure data center. You can also use this service to transfer data from Azure storage to hard disk drives and ship to your on-premises site. This service is suitable in situations where you want to transfer several terabytes (TB) of data to or from Azure, but uploading or downloading over the network is infeasible due to limited bandwidth or high network costs.

The service requires that hard disk drives be BitLocker encrypted for the security of your data. The service supports both the Classic and Azure Resource Manager storage accounts (standard and cool tier) present in all the regions of Public Azure.


Shipping
3. The customer ships the hard drives to the data center.
4. The carrier delivers the hard drives to the data center.
1. The customer prepares the hard drives using the Import/ Export Client Tool, and encrypts the drive with BitLocker.
2. The customer creates an import job using the Azure Portal.
Creating
5. The hard drives are processed at the data center.
6. The data is copied from the hard drives to the storage account.
Transferring
8. The hard drives are shipped back to the customer.
7. The hard drives are packaged for return shipping.
Complete
Packaging
Job state

APHIS has an existing Azure subscription which should have one or more storage accounts to take advantage of the Import/Export service. Each job may be used to transfer data to or from only one storage account. In other words, a single import/export job cannot span across multiple storage accounts.

APHIS can use Azure Import/Export service to copy data to Block blobs or Page blobs or Files. Conversely, you can only export Block blobs, Page blobs or Append blobs from Azure storage using this service. The service does not support export of Azure files and can only import files into Azure storage.

Azure does not employ mass migration devices like AWS Snowball, making large-scale data ingestion very difficult given the Import job can span only ten drives at a time. There would be an added cost of procuring drives for export to Azure.

The amount of time it takes to process an import/export job varies depending on different factors such as shipping time, job type, type and size of the data being copied, and the size of the disks provided. The Import/Export service does not have an SLA but after the disks are received the service strives to complete the copy in 7 to 10 days. You can use the REST API to track the job progress more closely. There is a percent complete parameter in the List Jobs operation which indicates copy progress. Reach out to AWS for an estimate to complete a time-critical import/export job.

3 Summary of Recommendations

CNSS suggests both near and long-term solutions that APHIS could leverage to get to the desired state where scientific data and processing are segregated from IT assets to optimize performance of both systems.

High-Performance Computing (HPC)

- APHIS needs to continue to work with ARS in overcoming the inter-agency hurdles to leverage the SCINet environment. Service Level Agreements and Memorandum's of Understanding need to be generated which will direct APHIS in how SCINet resources are accessed and what happens when situations occur such as National emergencies and exceeding capacity. Ames, IA and Ft. Collins, CO are colocation sites that are already enjoying some of the benefits of SCINet. Extending PPQ building 580 and adding SCINet points of presence (POP) in Riverdale is the next logical step in joining with SCINet.
- APHIS could leverage Azure Stack as an on-premises resource to augment the Microsoft hosted Azure solution. This hardware would arrive pre-configured and would be installed by APHIS's vendor, CGI, at no initial cost. Reoccurring costs will be charged for the platform resources used. The Azure Stack will allow APHIS to manage costs in an on-premises solution, to integrate more seamlessly with the Microsoft hosted Azure instance, and to exercise control of the environment through a single pane of glass.
- By segregating scientific big data and processing into SCINet, APHIS can better manage and consolidate all other data needs internally through Azure Stack. By enabling remote site based VDI outside of the data hubs, program personnel are afforded access to centralized resources and large data connections as if they were located in those locations. Eventually APHIS could join its scientific instrumentation to SCINet and start producing data sets directly in the scientific computing environment.

Big data and metadata storage and management

- There is a need to think about Big Data as many of the seemingly small data sets around the agency all seem to be growing and not shrinking. In many cases, programs have resorted to all sorts of media and storage alternatives that are out-of-band of the Enterprise Storage Strategy. Data exposure alone shows the need for a strategy to centralize scientific data and allow controlled growth, continuity, and security.
- While SCINet is the long-term goal with some short-term realization in Ames, IA and Ft. Collins, CO, Enterprise storage is going to play a large role in the centralization and consolidation of data around APHIS and enable programs to store everything centrally. Prior to this happening APHIS needs to replace their aging Oracle Pillar Axioms with the proposed Nimble C5K 310 TB Storage Arrays.
- While consolidation is underway, programs located outside Ames, IA, Ft. Collins, CO and Riverdale, MD need to start consolidating their mobile storage solutions into temporary storage systems such as the program recommended solution WD My Cloud PR4100, which scales up to 40TB. Mobile storage allows program locations targeted for or experiencing server consolidations to have a central local storage solution that could be phased out once the new centralized storage infrastructure is online. The point is to avoid profound program impact from potentially catastrophic failure. Once the storage transformation is complete these devices could be removed or used as onsite storage and staging data platforms for replication to the larger solution.
- Further discussion is needed on what can be stored on SCINet and whether or not APHIS needs to establish some storage enclave to store and deal with ever growing data sets.
- Implementing FIM and metadata options will ensure data integrity and accessibility both within APHIS and among collaborators.

High-speed communication for data sharing and collaboration

- SCINet affords the internal and external communications bandwidth, and a collaborative ecosystem that will enable APHIS to expand its global animal health leadership.
- In Ames, IA maneuvering data is simple because of the Ceres and storage platforms stored locally. Switches already exist that connect directly to the SCINet environment. When APHIS puts workstations in the labs and the Café, Scientists will then be able to migrate data into SCINet.
- Once storage has been consolidated, APHIS can start rolling out VDI to the program sites located outside of Ames, IA, Ft. Collins, CO and Riverdale, MD. This would allow the programs outside of these locations that are having server technologies consolidated, access to their data as if they were sitting in one of the hub locations. This could include access to SCINet in the distant future once the firewall between APHIS and SCINet is in place. A bridge will make SCINet services accessible to the masses without the need for expensive SCINet or Internet 2 POP's, which would be cost prohibitive in locations with fewer than 20 personnel.
- Until high-speed networks are available to more sites and scientific data is centralized, APHIS can leverage VDI, WD My Cloud PR4100 and Snowball type devices with more comprehensive backup and security measures implemented.

Effective Scientific IT support resources

- VRSC provides ARS Scientists with research IT support and specialized subject matter support such as informatics and GIS. Support includes installation and optimization of scientific software, helping with user questions, and providing tutorials and regular user training.
- APHIS needs to begin now to cultivate a parallel capacity to assist with sharing SCINet resources, provide guidance to APHIS Scientists to prepare for the HPC environment, and found the discipline within the organization.

Permanent big data leadership

- In order to develop effective Scientific IT policies, an acquisition strategy, and a detailed plan for optimizing, consolidating, and migrating systems APHIS needs to instantiate a separate SIT branch that is parallel to the current Enterprise IT organization. The SIT organization should begin now to establish an open culture that encompasses vertical and horizontal, central and decentralized points of view.
- Job instructions, official memos, policy statements, procedures, manuals, and similar organization wide artifacts must flow down from the top based on open and honest input from lower levels to management ensures both buy-in and common understanding.
- Coordination and integration of diverse functions is best served by horizontal and diagonal communication institutionalized through collaboration groups, cross-pollination of ideas, results, and methods, matrix structures for projects, and similar cooperative approaches.

Program Level Requirements

- APHIS programs have some unique needs, and CNSS has researched and recommended solutions that would mitigate them at the program level. Program-sponsored solutions have been evaluated with the criteria that individual programs implement these separate from APHIS IT initiatives as self-contained strategies.
- Recommended program level strategies include VDI, workstation upgrades, HPCs, FileMaker Pro, and numerous near-term options.

Access to Open Source Tools

- Adoption of a SIT environment, collaborating with ARS to move scientific research to SCINet, and adoption of AWS and AWS Direct Connect, and providing localized solutions as needed, will provide Scientists with access to numerous open source tools.

In addition, APHIS needs to manage and operate their existing applications in a modern way, while also taking advantage of the cloud model whenever possible. This transformation shifts APHIS from a traditional enterprise model to a hybrid-cloud model where systems and applications should be deployed simply as infrastructure code moving into a DevOps Model. This approach simplifies complex deployments and frees up Enterprise IT to focus on other tasks to better care for the environment. This scheme would also empower programs or developers with limited knowledge to deploy power infrastructures in the cloud or on premise from a service catalog established by Enterprise IT for each program.

The DevOps and Infrastructure as Code approach works whether prospective infrastructure resides on premise or in any cloud. With a few clicks of a button, scripts would do all the heavy lifting for rapid prototype deployment and modeling purposes. Code can be ported to a more sustainable environment once proof of concept is established.

These recommended solutions coupled with program suggested solutions can help APHIS and their programs realize a simpler, more integrated environment where tools for APHIS program personnel are made available to empower their work and scientific research.

4 Rough Order of Magnitude (ROM)

APHIS Enterprise IT Solutions

AWS

In this section, we have included some cost estimates and the related assumptions.

Estimates

- USDA HPC estimate On Demand in GovCloud
- USDA HPC estimate Partial Upfront Reserved Instance (RI) in GovCloud
- AWS HPC Reference Architecture

Assumptions

- Very rough initial estimates
- US GovCloud region – APHIS would like to start with the most restrictive environment, then potentially consider less restrictive regions later (US East/West)

Storage

- 1PB of storage in S3 Standard
  - Does not include expected increases of ~ 50 – 245% per year as this is too much of a range to calculate at this time.
- The storage of the 1PB of data with 10 million PUT/COPY/LIST Requests and GET requests
  - ***$30,242/month***

Snowball – Data migration Services

- To ingest 1 PB of data, thirteen (13) 80 TB Snowball units may be a one-time charge
  - ***$250.00*** per unit
  - This include 7 days (4 for shipping and 3 for utilization) of utilization
  - Ten units of 80TB (1,040 TB total space) Snowball for a single data ingest: ***$3,250.00***

HPC Cluster

- Initial HPC cluster of 4 instances of C4.8xlarge (36 cores each) based on APHIS stating each on premise server has 80 cpu/server
  - 36 cores each (144 total)
    - 60 GB memory each
    - 10 Gbps network each
  - 100 GB EBS volumes (400 GB total)
    - Provisioned IOPS SSD
      - 20,000 IOPS (320 MBs/sec)
  - EBS snapshot per day - 10% size of each volume
  - Egress costs are not included as this is an unknown at this time.
  - Used Partial Upfront Reserved Instances (RI) reduces hourly compute costs, but adds a one-time RI payment and no ongoing monthly costs.
  - Includes Enterprise level support
    - On-Time Payment: ***$16,924.00***
    - Monthly Costs: ***$8,154.65/mo***
    - AWS Enterprise Support ***$15,000/mo***
    - Total: ***$22,413.31/mo + one -time $16,924.00***

AWS Waiver

As mentioned previously, CNSS discovered an Egress charges waiver is in place for research and academic institutions that can provide a maximum discount of ***15%*** of the total monthly spending on AWS services that APHIS could apply either by being a research institution or by associating with academic institutions. APHIS would need to apply for the waiver through AWS.

AWS requirements for waiver of Egress charges:

- Work in academic or research institutions.
- Run any research workloads or academic workloads. However, a few data-egress-as-a-service type applications are not allowed under this program, such as Massively Online Open Courseware (MOOC), media streaming services, and commercial, non-academic web hosting (web hosting that is part of the normal workings of a university is allowed, like departmental websites or enterprise workloads).
- Route at least 80% of their Data Egress out of the AWS Cloud through an approved National Research and Education (NREN) network, such as Internet2, Energy Sciences (ESnet), GÉANT, Janet, Singapore Advanced Research and Education Network (SingAREN), Security Innovations Network (SINET), Australia's Academic and Research Network (AARNet), and CANARIE. Most research institutions use these government-funded, dedicated networks to connect to AWS while realizing higher network performance, better bandwidth, and stability.
- Use institutional e-mail addresses for AWS accounts.
- Work in an approved AWS Region.

Azure Cost Estimates and Assumptions

Assumptions

- Egress costs are not applicable due to Azure Stack being integrated into existing network.

Storage

- 1PB of storage in Azure Block Blob Storage
    - Does not include expected increases of ~ 50 – 245% per year as this is too much of a range to calculate at this time. The storage of the 1 PB of data is approx.
    - 10 million IOPS/mo
    - ***$44,409.85/month.***

Azure VDI

- This option is added as a cost comparison to what an on-premises solution would cost if the data was moved into the Azure Cloud.
- Initial estimate of 3000 Virtual Desktop (Azure Virtual Machines Windows VM) to deploy Azure Stack. Estimated cost based on 10 core hours per day for a month. The Virtual Desktop for 300 machines is approx.
    - ***$31,800/mo*** (base cost $0.53/10 core hours).

Azure HPC Cluster

- Initial HPC cluster of 8 instances of H16r (16 cores each) based on APHIS stating each on premise server has 80 cpu/server
    - 16 cores each (128 total)
        - 112 GB memory each
        - 10 Gbps network
    - 50 GB EBS volumes (400 GB total)
    - Egress costs are not included as APHIS already has a 1 Gbps connection to the Azure Cloud
    - Monthly Costs: ***$20,197.12/mo***

On-premises VMWare VDI

VMWare Horizon Enterprise vCAN Licensing

- APHIS has brokered a contract with VMWare to utilize vCloud Air Network (vCAN) licensing usually reserved for service providers.
  - This cost comparison is provided to show what it would take to virtualize 3,000 virtual desktops to access the data that would be centralized in Ft. Collins, CO and Riverdale, MD.
- The VMware vCAN offers service provider partners the option to "rent" VMware software licenses on a monthly Pay-as-You-Go (PAYG) basis, which in turn allows service providers to provide VMware hosting products and create customized hosted infrastructure solutions.
- APHIS is entitled to $1,000.00/mo in credits for DEMO/Test purposes
  - Any VDI used for testing or Proof of Concept (POC) would fall into this category

Assumptions

- APHIS will be using vCAN licensing which is a PAYG licensing model reported monthly to VMWare
- APHIS has ~7000 employees with ~4100 employees in Ames, IA, Ft. Collins, CO and Riverdale, MD
  - This model assumes that ~3000 VDI will be needed on the vCAN licensing model

Pricing

- Under vCAN licensing the VM configured will be:
  - 4 virtual central processing unit (vcpu) per VM
  - 8 GB memory per VM
- The average point cost for each VM would be ***20 points (pts)***
  - Total of ***60,000 pts***
- APHIS pays only ***$0.54*** per point when between 30,000 and 100,000 points are reported
  - 360 - 1800 $0.94
  - 1800 - 3600 $0.82
  - 3600 - 10800 $0.71
  - 10800 - 18000 $0.65
  - 18000 - 30000 $0.59
  - 30000 - 100000 $0.54
  - 100000 - 250000 $0.50
- Total cost for ~3,000 VDI would be ***~$32,400/mo***
  - This price would fluctuate each month depending on how many VDI were reported on

APHIS Program Centric Solutions

Remote office on-premises solutions

Western Digital

Western Digital My Cloud Pro Series PR4100 with a 40TB capacity

- Cost for a single unit of WD My Cloud PR4100 is ***$2,199.99.***
- Estimating ~20 units of the WD My Cloud
  - An approximate one-time cost of ***$43,999.80***

GIS Desktop based workstations

Dell Precision 5720 All-in-One

- Cost for a single unit is ***$4,183.33.***
  - Estimated 10 workstations needed: ***$41,830.33***

- An APHIS study would need to be conducted of the GIS workers on whether or not they could utilize the Esri cloud services or if their current workstations are sufficient for their needs.

FileMaker Pro – Application front-end for datasets

While not a BI solution, FileMaker Pro allows PPQ to develop their own interface for existing datasets. ePermits would remain in its existing as-is state and FileMaker Pro would give BRS and PPQ the ability to create their own front-ends with no programming to take advantage of custom visualization of their datasets.

FileMaker Pro 16 Advanced

- Integrates with Tableau via a web-connector
- Can be stood-up in AWS Marketplace
    - Bring Your Own (BYO) license
- Identified for the PPQ PGA Message Set
    - Estimates 10 user license at ***$1,428.00/per year***

File Integrity Monitoring

Log & Event Manager

Log management software for security, compliance, and troubleshooting

Key Features

- Fast and easy compliance reporting
- Real-time event correlation
- Real-time remediation
- Advanced search and forensic analysis
- File integrity monitoring
- USB device monitoring

Starts at ***$4,495*** one-time charge per license

Business Intelligence Software

CNSS has compared and made recommendations from section 2.0 above that encompass long-term, short-term and program specific solutions. The below BI packages are provided as a comparison in costs. The pricing is forecast on a 300-user license or by a specific program's data volume.

Tableau

Tableau Desktop Professional Edition

- Cost for a single unit of Tableau Desktop Professional is ***$70 Per User/mo***.
    - The costs are billed on an annual basis.
- Estimating 300 users the approx. annual cost is ***$252,000 /yr.***

Tableau Server On-premises/Public Cloud

- Could be utilized for one-off editions
- Cost for a single server of Tableau Server is ***$35 per User/***mo.
    - The costs are billed on an annual basis.
- Estimating 300 users the approx. annual cost is ***$126,000 /yr.***

Microsoft PowerBi

Identified by the PPQ PGA Message Set as they are using the Power BI Desktop today.

- Power BI Desktop

  - Free for an individual user
- Power BI Pro
  - $9.99 per month per user
    - 300 Users ***~$3,000.00/mo***.
- Power BI Premium
  - 300 pro users, 150 frequent users and 330 occasional users (500 approximate users)
    - Power BI Pro licenses
      - ~$3,000/mo
    - Power BI Premium: 1 x App Services P1 Node
      - ~$300.00/mo.
  - Total: ***~$3,300.00/mo***

RapidMiner

- The RapidMiner Education Program provides free RapidMiner product licenses for academic usage to students, professors and researchers.
  - It's feasible to think that APHIS could procure a free license for the purposes of research as a research institution

Pricing

	FREE	SMALL	MEDIUM	LARGE
# Data Rows	10,000	100,000	1,000,000	Unlimited
# Logical Processors	1	2	4	Unlimited
Performance Improvement		2x	4x	10x+
Background Process Execution				X
	$0.00	$2,500/yr	$5,000/yr	$10,000/yr

- Given the size of the PPQ PGA Message Set is 15 million+ records and growing, the program would have to procure a ***$10,000/yr*** license

Appendix A: Documents Reviewed

During the course of the engagement with APHIS, CNSS reviewed the following documents provided by APHIS and affiliated vendors.

- CTO
  - ITD Management - Azure Cloud Aug 2017 v0.1.pdf
  - DCOI Field Site Inventory Consolidation Spreadsheet 03-28-17.xlsx
  - APHIS_GSSHardwareSpecs.xlsx
  - APHIS Server%2c SAN and Network Architecture.pdf
  - All Workstations Win10 Compat.xlsx
  - m_16_19_1.pdf
- VS
  - USDA APHIS VS STAS: Scientific Computing Design and Implementation
  - USDA_Network_Detail_v0.3.pdf
  - ARS SCINet diagram - 20150210.pdf
  - APHIS_VS_STAS_BioTeam_Assessment_Final_Report.pdf
  - 072017 VS CEAH Response to Scientific Computing Assessment Template.doc
  - 071917 VS Ames response - Scientific Computing Assessment Template.doc
  - 2017-07-26.pptx
  - 7-21-17 VS FADDL Scientific Computing Assessment revised questions.docx
  - BDSC Alternatives Analysis - Final.pdf
  - BDSC Alt Analysis Cost - Detailed with Comments v2.xlsx
  - Ames-vSphere6.0_UpgradeSchema_11-7-16.pdf
- BRS
  - Scientific Computing Assessment Template 071417-brs2.doc
  - Scientific Computing Assessment Template 071417-brs.doc
  - Initial Response.docx
- AC
  - Big Data Project Questions (1).docx
- WS
  - APHIS Summary of Findings - Submittal Draft 08142017 BEW.DOCX
- AWS
  - AWS Response to APHIS HPC Market Research.pdf
  - USDA NetBond Use Case Diagrams%5b4%5d.pdf
  - NetBond White Paper.pdf
  - NetBond Cost Calculator v2%5b2%5d.xlsx
- Azure
  - Getting_Started_With_Azure_Resource_Manager_white_paper_EN_US.pdf
  - Azure_Stack_an_Extension_of_Azure_EN_US %281%29.pdf
  - Azure Gov Cloud Service Catalog Price Sheet September 2017
  - Azure Gov ARM VM Size- Cost Captures.xlsx

Appendix B: APHIS Summary of Findings 09/11/17

CHEROKEE NATION
System Solutions

APHIS Big Data Assessment

Summary of Findings

October 19, 2017

Submitted by:
Cherokee Nation System Solutions, LLC
Trayci Koppie, Technical Project Manager
C: (703) 898-5942 | email: Trayci.Koppie@cnt-fc.com
777 W. Cherokee St. Catoosa, OK 74015 |
4803 Innovation Drive, Suite 3 Fort Collins, CO 80525|

Table of Contents

Project Summary

The Cherokee Nation System Solution (CNSS) summary of findings provides the data collected through the scientific computing assessment and the location interviews with United States Department of Agriculture (USDA), Animal and Plant Health Inspection Service (APHIS) management, scientists and IT personnel. This information includes organizational information, day-to-day processes, products utilized and environmental data. The summary of findings is the collection of all information requested from the six programs supporting APHIS, Veterinary Services (VS), Plant Protection and Quarantine (PPQ), Animal Care (AC), Wildlife Services (WS) and Biotechnology Regulatory Services (BRS). It also includes ancillary findings from Agriculture Research Service (ARS) as a partner agency that shares similar IT requirements and mission needs that may provide efficiencies by consolidating infrastructures.

Big Data Assessment Background

Basics of Big Data – What is Big Data?

Big data is a term for datasets that are so large or complex that traditional data processing capacity and application software is inadequate to deal with them.

APHIS uses big data for genomics; population, disease and geospatial modeling; analysis of surveillance data; molecular analyses; and risk assessments. For example, think about the recent comparisons on the genomic sequencing of avian influenza viruses to compare the low and high-pathogenic sequences. We continue to use and analyze these data as we conduct our epidemiological analyses. How and when did the virus get introduced into commercial poultry flocks and where did it drift or shift along the way?

Think about the additional possible uses:

- As the 340 rule goes through, applicants may be inclined to provide genetic sequencing along with their applications for permits or for de-regulation. We could analyze the sequences to model and predict functions.
- It would be helpful for wildlife services to confirm – in real time – that an animal they have captured or euthanized is the same one that preyed upon a flock of sheep.
- Being able to provide genetic evidence to link an outbreak of a plant disease to an illegally imported cutting could vastly improve enforcement capabilities.

Datacenter Optimization Initiative (DCOI)

APHIS, like all other USDA entities, fall under the DCOI initiative by the Whitehouse to reduce their overall server and data center footprint as a requirement under Federal Information Technology Acquisition Reform Act (FITARA) ratified in 2015.

The bill outlined here (https://datacenters.cio.gov/assets/pdf/Significant%20Expansion%20Definition%20White%20Paper_v2%203.pdf) has far reaching effects such as requiring agencies to track costs resulting from implementation of the Initiative within the agency and submit an annual report on such costs to the FCIO.

The bill expresses the sense of Congress that transition to cloud computing offers significant potential benefits for the implementation of federal IT projects. The bill also permits CIOs to establish cloud service working capital funds.

As a result of DCOI, APHIS is now looking to consolidate a majority of the server technology in offices with less than 20 personnel to either Ft. Collins, CO, Riverdale, MD or the Microsoft Azure Cloud.

Table 1 Ref: APHIS DCOI Field Site Inventory Consolidation

Program	Total Sites	Total Servers	Servers Kept	Servers Removed	Total Storage (GB)
AC	1	1	1	0	931
IS	28	34	30	4	67,016
PPQ	103	127	28	99	534,208
WS	27	30	3	27	172,179
VS	21	22	4	18	78,347
Shared Sites	24	37	14	23	144,870
Total	**204**	**251**	**80**	171	**997,551**
60% Reduction	150.6				
Remaining	**100.4**				

With an overall reduction of 68% of the APHIS compute environment and the elimination of 213TB of organic storage, APHIS will need to either consolidate on centralized Enterprise Storage or move their Scientific Data into the cloud.

Microsoft Azure APHIS

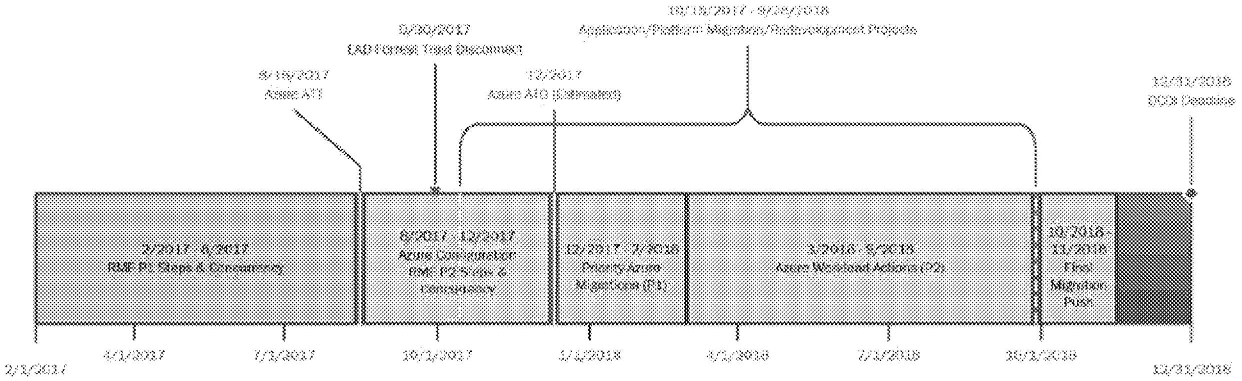
Background

APHIS has adopted Microsoft Azure as its official cloud platform in order to help meet DCOI compliance and enabling resources for the APHIS programs to take advantage of its ability to rapidly expand capacity for existing Information Systems that would be ported. APHIS is not targeting Azure for porting virtual machines, but to take advantage of its unique platform abilities that will allow on-demand growth.

Microsoft Azure offers the ability for custom applications to be grown using native services vice virtual images allowing for more dynamic scaling than what the confines of standard virtual machines offer.

APHIS has achieved RMF Phase I and has obtained an Authorization-To-Test (ATT) in the Microsoft Azure Cloud allowing for the initial testing of workloads in the cloud and refining migration plans. September is the planned pilot where APHIS will officially test the Azure platform.

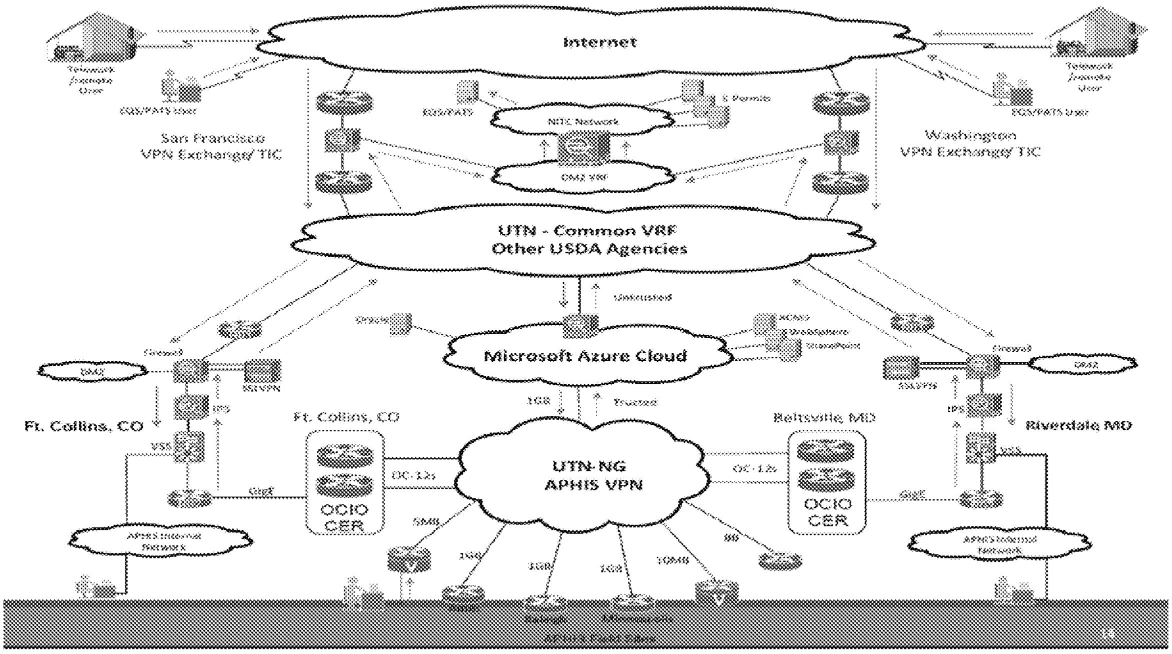
Table 2 Azure Timeline



APHIS Azure Architecture

APHIS has already had a converged architecture via Azure ExpressRoute via a 1Gbps connection from the APHIS UTN-NG VPN cloud.

Table 3 APHIS USDA converged network architecture



Veterinary Services (VS) Summary of Findings

1.0 Background

The National Veterinary Services Laboratories (NVSL) safeguard U.S. animal health and contribute to public health by ensuring that timely and accurate laboratory support is provided by their nationwide animal-health diagnostic system. NVSL staff accomplishes this through:

- Providing diagnostic services, reagents, and training in world-class facilities
- Responding to animal health emergencies
- Taking an active role in managing the National Animal Health Laboratory Network (NAHLN)
- Serving as an international reference laboratory
- Maintaining a well-trained and responsive staff

The Center for Veterinary Biologics (CVB) ensures that veterinary biologics available for the diagnosis, prevention and treatment of animal diseases are not worthless, dangerous, contaminated or harmful.

- Provides expert and statistical analysis of data to support safety and efficacy of products.
- Performs laboratory testing on biological products and components.
- Collects and analyzes data to improve laboratory testing, reduce animal use and improve pharmacovigilance.

The Center for Epidemiology and Animal Health (CEAH) is one of the science centers within the USDA, APHIS, VS, Science, Technology, and Analysis Services (STAS) that:

- Promotes and safeguards U.S. agriculture by providing timely and accurate information and analysis about animal health and veterinary public health.
- Explores and analyzes animal health and related agricultural issues to facilitate informed decision making in government and industry.
- Provides a collaborating center for surveillance, risk assessment, and epidemiologic modeling with the World Organization for Animal Health (OIE). In this role, we interact with collaborating centers domestically and internationally and participate in training activities.
- Staffs a multidisciplinary team, providing scientifically sound and statistically valid information and tools to help policymakers reach critical decisions regarding animal health issues.

The NVSL Foreign Animal Disease Diagnostic Laboratory (FADDL) is part of the only facility in the United States where many infectious foreign animal disease (FAD) agents are studied. It is located on Plum Island, which is 1.5 miles from the northeastern end of Long Island, New York. Scientists at the FADDL are devoted to diagnosing foreign diseases of animals. They partner with scientists of the Department of Homeland Security (DHS) and ARS, also located on Plum Island, in foreign animal disease research.

Additionally, the FADDL is the custodian of the North American Foot-and-Mouth Disease (FMD) Antigen Bank. The Bank stores concentrated FMD antigen that can be formulated into vaccines if an FMD introduction occurs. The Bank is co-owned by Canada, Mexico, and the United States. Personnel working in the vaccine bank are responsible for performing safety and potency testing of new antigen lots of FMD vaccine, and periodically testing the quality of stored antigen.

2.0 'As-Is' Environment

The VS location in Ames, IA contains an active 100 GB connection to Internet2 (see 2.0.1) that is segregated from the APHIS Enterprise Network connected to the USDA Unified Telecommunications Network (UTN). Internet2 allows for fast transport between the Amazon Web Services (AWS) cut out into the ARS SCINet Network (see 2.0.2). The APHIS building in Ft. Collins, CO has a subsequent 10 GB connection to ARS SCINet and is used primarily as a backup site to Ames, IA. There is no Internet2 transport between the APHIS network into the ARS SCINet Network (see 2.0.2). There is a workstation planned to be connected in the café at Ames, IA that does allow for data to be transferred to SCINet. Due to time and resource constraints, it has not yet been a priority to install, thus resulting in no public aperture to upload large data-sets to SCINet. This gives Scientists at Ames, IA the ability to have data transferred via portable storage media. There is a desire for access between the APHIS network and the Internet2 (SCINet) environment but how this impacts the original "air gap" stance of the SCINet design will determine if this is possible. There is a desire to install a local firewall in Ames between SCINet and the APHIS network. While the ARS SCINet has an existing firewall that segregates the untrusted network to the accredited network, it is agreed that APHIS should install their own firewall to ensure that their unique security concerns can be addressed between the APHIS demarcation and ARS SCINet border.

2.0.1 SCINet powered by Internet2 – The History

Internet2 is a member-owned advanced technology community founded by leading higher education institutions in 1996. Internet2 provides a collaborative environment where research and education organizations can solve common technology challenges and develop innovative solutions in support of their educational, research and community service missions.

Internet2 is the nation's largest and fastest coast-to-coast research and education network and comprises:

- 317 U.S. institutions of higher education
- 81 leading corporations
- 64 affiliate and federal affiliate members
- 43 regional and state education networks
- More than 65 national research and education networking partners representing over 100 countries

The Internet2 community touches nearly every major innovation that defines our modern digital lives—and continues to define "what's next."

While Internet 2 may be accessible at all Points of Presence (POPs), SCINet remains limited to Internet2 users unless explicitly given access. This means that extract work at an Internet2 POP will have to be done to attain access to SCINet resources.

2.0.2 SCINet Findings

CNSS conducted interviews with Scott Farris from VS and Victor Unruh from ARS in regards to the current "As-Is" SCINet infrastructure. While at the Ames, IA facility (a joint ARS and APHIS facility), Scott Farris took CNSS on a tour of the server room to see the CERES (The Goddess of Agriculture) HPC and its associated storage system.

It was confirmed that there is a point of presence from Ames, IA and Ft. Collins, CO on SCINet running at 10Gbps overall (some sections of SCINet have 100Mbps connectivity, but not end to end in many cases which limit the overall speed) which is ample for the current needs of those facilities. This network is completely parallel to the existing APHIS network with only a single connection point via the internet and through an Amazon Web Service (AWS) VPN tunnel protected by a firewall. This would allow scientists to run remote session via SSH directly to CERES or transfer datasets, programs or any other data needed to support processing in series.

In conversations with Marco Munoz and Scott Farris, it was discussed that APHIS may desire to implement its own firewall to ensure that the APHIS required security posture is met. ARS would keep their firewall "As-Is" and APHIS would add one of their own in front of the APHIS demarcation point into the SCINet network.

In talking with Victor Unruh from ARS in Ft. Collins, it was found that the CERES HPC Cluster is currently running at 70 – 80% utilization but could be scaled rapidly to meet the needs of APHIS.

2.0.3 High-Level SCINet Network Diagram

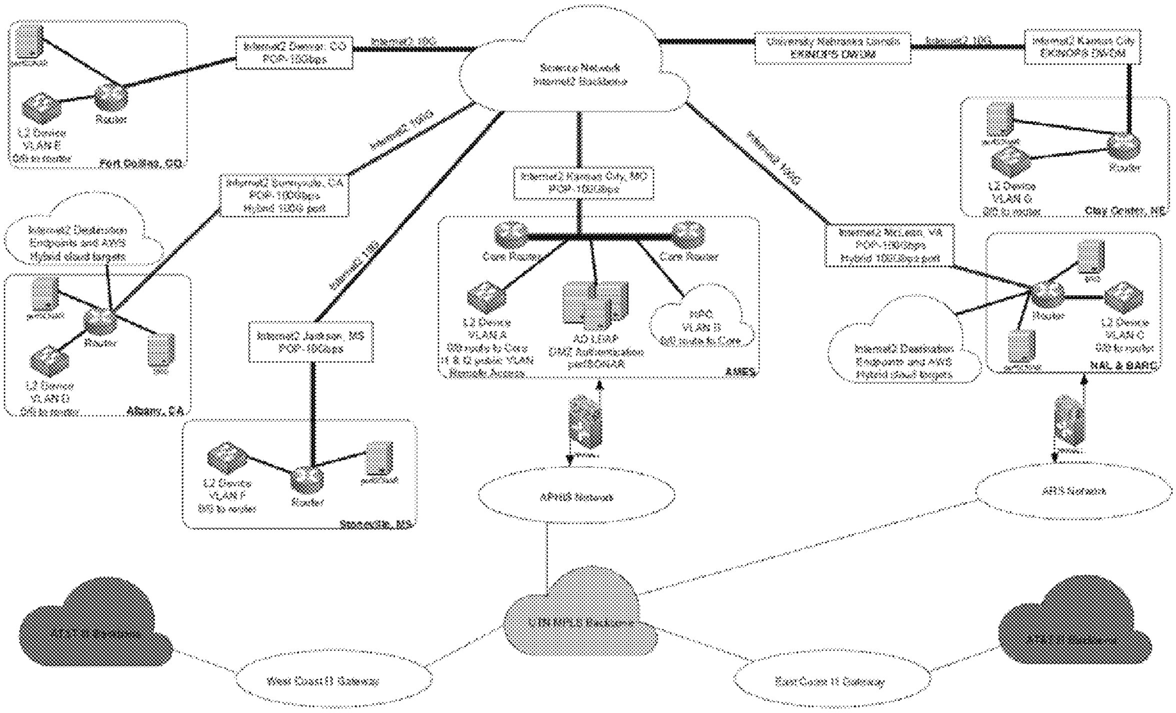


Figure 1 - Depicts the current ARS SCINet network.

2.1 Storage

2.1.1 BIOINFO Storage

The VS Bioinfo drive is the shared storage that NVSL/CVB uses to converge the PC/MAC environment with the Red Hat Enterprise Linux Servers on the APHIS Enterprise Infrastructure (AEI). This environment is not currently replicated between Ames, IA and Ft. Collins, CO via the APHIS Enterprise Network. Current data size constraints prevent serving as an offsite backup and Continuity of Operations (COOP) site for the Scientific Data. The data in Ames, IA instead is backed up to tape and shipped to Ft. Collins, CO. Plans for a full replication of the data from this location are planned when the storage solution is modernized.

This multi-site topology allows for sharing of datasets between sites. The storage is located on the APHIS Enterprise storage. VS would like to migrate the data from the enterprise network/Storage to the Seagate CS9000 Luster Parallel File System environment on SCINet. This is a desire, not an existing state for the VS data. The Seagate File System allows for a maximum speed of 63Gbps per rack and up to 1Tbps of throughput overall into the storage array. However, the current footprint is a fraction of the current network speed potential but has plenty of room to grow.

The current storage utilization is approximately 76TB of Raw Scientific Data located on the standard Pillar Enterprise Storage. This could not be completely quantified since each scientist also stores scientific data in their Enterprise Share causing data duplication and lack of centralization of all scientific data. There is an evident lack of storage available in the enterprise storage system which is growing at a rate of 235% per year on the average. In addition, all of the storage capabilities described up to now are not being used for modeling, which has additional storage requirements.

The Pillar storage medium is near the end of life and at maximum capacity. Although the Pillar is being replaced with another enterprise storage solution, decisions will need to be made as to where the future scientific data will be stored. All scientific data needs to be centralized in a scientific network to reduce the burdens of the Enterprise Storage platform and reduce data duplication across the networks.

The Antimicrobial Resistance (AMR) Program has a growing need for storage, although not yet quantified. This program is expected to rely heavily on Big Data with multi-year storage needs. This data will have to conform to Confidential Information Protection and Statistical Efficiency Act (CIPSEA) which impacts a majority of the AMR program. CEAH highlights great urgency around the emerging AMR Program. It is expected that this program will rely heavily on scientific computing and Big Data. Key computing needs will be:

- Bioinformatics platforms to support analysis of and potential storage of genetics data potential integration with NVSL or another lab external to VS.
- Data storage capabilities for studies that span over multi year.
- A data enclave to support collaborative computing with Federal, State, and Academic partners for scientific computing and advanced analytics.
- A significant portion of the AMR Program's work will fall under the control of the Confidential Information Protection and Statistical Efficiency Act (CIPSEA) which creates unique data security challenges and not generally able to share environments with other computing environments not using common security controls.

2.1.2 Enterprise and Scientific Data Comingling

VS lacks a location for all scientific data to be worked on. Instead scientific data is co-mingled with enterprise data dispersed on centralized storage, portable media and individual workstations. Scientists would be better served having their data in a single location to accommodate their needs for modeling. Scientific data is a small fraction of the data represented in the enterprise storage platforms.

While this is a good location for disaster recovery or COOP purposes, the enormity of these datasets has caused Enterprise IT to require the scientists at Ft. Collins to remove this data from the enterprise systems forcing them on to memory sticks, hard drives or other network shares they can find to store their data. Often these systems aren't part of the Enterprise Disaster Recovery Strategy. This could lead to catastrophic loss of invaluable scientific data and leaves the organization open to infection of malware using uncontrolled storage devices within the APHIS Enterprise Network. This is a major cause for concern and needs to be addressed as soon as possible.

2.1.3 Diversity of data sources

VS uses a variety of instrumentation for the acquisition of scientific data such as gene sequencing. A majority of these instruments are located on the network and have the ability to start amassing into large datasets. VS also acquires data from external sources through manual curation, which is one of the most time-consuming activities. Often these datasets have flaws in them that have to be fixed prior to the scientists or analysts being able to utilize them in models for analytical purposes. Once these datasets are used, they typically have to be reconstituted in the same manner costing valuable time and creating a tedious process.

2.1.4 Metadata Based Indexing

One of the most challenging aspects of cataloging is determining what is important to keep in highly available fast throughput storage and what could be archived to a lower cost storage solution. This depends on knowing what type of data exists and information about it. This is where Metadata comes in as it describes what is in the dataset and is fully indexable by a variety of solutions.

VS in Ames, IA is looking into two suites that would assist with this issue:

- iRODS
- NIRVANA for Data Management (Is now being retired)

iRODS is an open source solution that would be cheaper in upfront costs but requires an extensive knowledge of the platform to get implemented in short order. NIRVANA would be quicker to set up but has a higher capital cost which could be comparable to the open source implementation costs. Both platforms run on any storage system and index your metadata into a Relational Database Management System.

Both platforms tag and allow for better use and prioritization of datasets. However, with the discontinuation of the NIRVANA platform, VS continues to search for alternatives to compare to iRODS.

Further exploration is warranted to better define what data should stay in the more expensive enterprise platforms and what can be archived off.

2.2 Security

2.1.1 External Datasets

VS has unique security needs depending on the type of data that it is receiving from outside sources (Educational Institutions, Labs, Farms, Private Industry, and Biologics Manufacturers, etc.). A majority of the information contains Personally Identifiable Information (PII) from external sources and has to be protected.

Datasets containing PII are kept on accredited APHIS Information Systems and Storage. Moving this data to another network with lower security requirements could be problematic since the datasets and PII can be integrated. Data minimization procedures to remove the PII altogether do not currently exist.

2.1.2 Enterprise Updates and Impact to Scientific Computing

Scientists can have long running processes on workstations connected to the APHIS network that an enterprise update may interrupt. They have been dealing with this issue by disconnecting the workstation from the Enterprise Network to run their processes and then reconnecting once the run has completed moving the dataset for further analysis.

2.1.3 Non-Compliant Instruments Connected to the Network

Issues with non-compliant Scientific Instrumentation connected to the network have been dealt with by Pseudo air-gapping it from the Enterprise Network through multi-homed network interfaces on compliant workstations. The scientists use Remote Desktop Protocol (RDP) to connect to the workstation that has the scientific instrument connected and they generate the dataset from the instrumentation then transfer it to either a network share or portable media for physical relocation of the data.

2.1.4 Confidential Information Protection and Statistical Efficiency Act (CIPSEA)

The Monitoring and Modeling (M&M) unit within VS CEAH in Ft. Collins, CO routinely works with data covered under CIPSEA. The confidentiality of this data is guaranteed by federal law and stringently managed by VS CEAH staff in the NAHMS data lab. Record level data is never openly released for any reason. This lab consists of an air gapped local area network consisting of a file server and four connected workstations within a limited access controlled physical space. Electronic data is moved into or out of the lab according to a strict protocol of clearance, tracking, and through the use of a password protected IronKey USB device. Only aggregated, statistically reviewed data is allowed out of the lab and then it is managed on the APHIS enterprise network for analysis. All data products associated with CIPSEA data undergo a strict review process before they are published and released out of the M&M unit.

2.3 Data Sharing and Transfer

2.3.1 External Data Interchange

Interviews with Scientists and IT Staff at both Ames, IA and Ft. Collins, CO has revealed that there is a challenge in getting large datasets sent between sites, teams or external organizational contacts. Typically, the data is manually loaded on some sort of portable storage (such as a memory stick or external hard drive) and sent via FedEx to external locations for collaboration. This often causes delays in datasets being analyzed or true collaboration due to the asynchronous nature of how the data is interchanged.

High-security data such as CIPSEA is currently interchanged by redacting the CIPSEA protected elements from the products. Once approved for release, the product is converted into a PDF file and is shared with the requestors through normal data interchange means (i.e. portable media or email).

They will need secure data enclaves that will enable them to collaborate with other groups using CIPSEA protected data because putting this level of data on uncontrolled systems is not an option for M&M.

2.4 Scientific Computing

2.4.1 Local Computing Resources

The VS location in Ames, IA has two local servers with the fairly high capacity to conduct a bulk of its computing needs. These systems have 64 AMD Processors, 512GB RAM and dual 800GB of SSD storage running Linux Red Hat Enterprise Linux.

The Ames, IA servers are not currently clustered and are used as individual stand-alone systems. The current setup does not allow for maximizing of local resources to speed intensive computational tasks. A lack of a scheduling system makes it more difficult for scientists and IT staff to coordinate the use of the localized resources to schedule multiple jobs and order of priority for each job with computational resources. SLURM Workload Manager is being used on CERES (HPC) in SCINet successfully to control jobs for SCINet contributors and could be used to maximize resources in the localized Red Hat Computing environment.

A third server currently located at Plum Island, NY has 44 Intel Processors, 756GB of RAM and 1.6TB of SSD storage available for local computational needs.

2.4.2 High-Performance Computing (HPC)

VS and ARS scientists in Ames, IA have been using the CERES HPC Cluster in SCINet for their large computational needs with good success. However, data transfer via portable media between the APHIS enterprise network and SCINet has proven to be very slow, often taking days to weeks to complete. Currently, FTP is being used for data interchange between the two environments via AWS as a current workaround as the APHIS and SCINet networks are not yet directly connected due to an accreditation issue.

This current setup is causing issues for pushing datasets up through this construct. For example, the latest challenge presented was that it takes a week for a 100GB dataset to be transferred from APHIS enterprise networks to SCINet via the AWS cut out. This is one of the main barriers to using this resource.

Applications or tools that are currently being used on workstations, laptops, and Red Hat Enterprise Linux server may have to retool to work in the HPC environment.

2.4.3 Ad-Hoc Computing

At both Ames, IA and Ft. Collins, CO locations, VS has reclaimed end of life workstations or laptops. These devices are utilized as an ad-hoc computing environment for the scientists to run their models. These environments are the most common since getting access to them is the easiest method despite processing runs taking a long period of time. It is the path of least resistance and therefore the typical solution. A short-term solution to this problem is needed to fill the gap while longer-term options are explored.

In Ft. Collins, CO, a small ad-hoc computer lab of 10 workstations has been established with no formal procedures for scheduling as a stop gap for the lack of localized resources.

2.5 Cloud

2.5.1 Enterprise Cloud

APHIS is currently undergoing a trial for Microsoft Azure Cloud in order to comply with the current Data Center Optimization Initiative (DCOI) as mandated by the USA CIO. This environment is conducive for more enterprise workloads but lacks a dimension for large scale scientific computing.

2.5.2 Scientific Cloud

VS in Ames, IA has started to look towards and has the authorization to use some limited Amazon Web Services (AWS) computing. The modular services model with Amazon would allow scientists to create bite sized workloads that can be scaled against a large computer infrastructure. Everything from high throughput processing to Peta Scale Storage and Machine Learning Algorithms would give scientists a large modular platform that can be used to meet the emerging needs in a modular fashion.

Manufacturers of scientific instrumentation such as the PACBIO and NextSeq Sequencers are taking advantage of the AWS cloud by offering on-demand access to their information systems in a private cloud setup for the initial processing followed by on premise analytics of the raw output. Scientists in VS have only been authorized to use the PACBIO AWS AMI's after over a year of review by APHIS. The inability to take timely advantage of the emerging cloud platforms is hampering efforts for scientists at VS to meet the needs of their workloads and demands being placed on them for faster analytical work, especially in a crisis scenario.

3.0 Miscellaneous Findings

3.0.1 Relocation of Plum Island

With the relocation of Plum Island, NY to Manhattan, KS on or about FY2023, there is an opportunity to take advantage of tying into an Internet2 PoP for further integration into the ARS SCINet network. This would allow for exponentially more processing than the single Linux Red Hat Server that is currently on-site.

Better data interchange could occur between Ames, IA, and Manhattan, KS reducing the need to FedEx external storage systems or perform slow transfers that take a week or more for datasets to be exchanged for collaboration purposes.

Moving FADDL from Plum Island, NY to Manhattan, KS depends on the implementation of the Select Agent Registration which will take place between 2021 and 2023 that may impact the near-term transition of the facility and use of the Internet 2 PoP as outlined in the NBAF Update (https://www.aphis.usda.gov/animal_health/downloads/sacah/2016/nbaf-lautner-kappes.pdf).

3.0.2 Open Source

Currently, VS in Ames, IA utilizes the following open source suites:

OPEN SOURCE SCIENTIFIC APPLICATIONS			
BWA	ABySS	Shell cmds	RAxML
GATK	Kraken	Git	t-coffee
Picard	kSNP	Perl	LaTeX
Samtools	Bamtools	Python	
IGV	Bbmap	R	
SPAdes	Newick utils	Local BLAST	

Using open source has been critical to the success of scientists and analysts alike. These tools allow them to cut down on manual processes or exchange a scientific model/process in a public manner for peer review or validation of work.

VS in Ames, IA currently hosts a GITHUB standard account and is looking to expand to an enterprise account for more flexible options.

Getting access to open source suites or creating repositories for code has been a painful process for VS in Ames, IA and other APHIS programs due to the lead time required gaining approval to allow even an evaluation of the software. Scientists have to use outside systems or wait for approval to find out if the tools meet their needs on APHIS datasets. There have been cases where it took months for approval and then in a day, it was found that the tool did not yield the proper results.

APHIS has underscored a need for a sandbox environment which will allow for rapid prototyping, testing and overall integration of these software suites prior to being put into an accredited APHIS system.

3.0.3 SCINet Inter-Agency Challenges

While there are challenges in using SCINet on the security and connectivity front, there seems to be a larger issue at hand. Processing time on CERES is based on a first-come-first-serve model and is prioritized via the Slurm Workload Manager.

VS has brought to light that they are involved in emergency events where it would be conceivable that they would need the highest priority in computational and storage resources for the duration of the emergency. APHIS policy makers need information from scientific computation as rapid as possible to ensure that timely decisions are made in complex situations. A current lack of a Memorandum of Understanding (MOU) between ARS and APHIS makes an elevating priority on CERES during an emergency difficult at best as it would require the suspension of on-going processing from other SCINet customers who may already have a higher negotiated priority.

Security posture on SCINet would possibly be a problem. SCINet would need to be accredited to store and process information at the appropriate level for APHIS data, which could be at a higher level than required by ARS. This could mean that a more secure enclave would have to be created from within SCINet or the mission assurance level of SCINet may need to be brought up to APHIS standards before data could reside in that enclave.

Failure to do so could cause a host of issues such as:

- Having to segregate PII or regulatory information from datasets
- Transfer of the data from APHIS to SCINet only for processing and then back to APHIS which would be tedious and time consuming
- Given the current speeds at which data can currently be interchanged with SCINet, processing would be more time consuming if the data did not reside in SCINet already

3.0.4 Relocation of CERES on SCINet a potential Issue

ARS is contemplating the move of CERES from the HCAH Campus in Ames, IA campus to the Iowa State University campus a few miles down the street. While this makes sense for ARS and academic involvement of the platform, it may cause complications moving forward with APHIS regulated data residing in a non-government enclave. The move has yet to be determined by ARS and they will continue contemplating what best would serve SCINet in the long run.

3.0.5 Geospatial Information Systems (GIS)

The bulk of GIS analytical work is done from the Ft. Collins, CO campus via desktop applications. A push by APHIS has been ongoing to move that platform into Esri's cloud platform to reduce costs. However, many sites are remote throughout the United States and the world and have expressed concerns that moving to a cloud-based platform may depend upon greater bandwidth requirements and computer performance may not be practical.

If the other APHIS agencies are similar to Wildlife Services, our GIS users are distributed in offices (many of which are remote) throughout the USA (and the world), so although cloud-based solutions might look appealing, the current state of connectivity and computer performance (without major upgrades) will likely make this an unreachable/impractical goal.

Most of the GIS data consumed by VS and APHIS are sources from externally facing systems or databases that come from a variety of sources such as state based databases. Metadata is downloaded and datasets are prepared in ArcGIS while the analytical works are completed with Aleryx and visualization handled by Tableau.

The total estimated sum of the data storage for GIS data is less than 1TB in total. Data is often not kept after analytical products are manufactured as the datasets themselves are derived from outside sources. This makes VS more of a data consumer rather than a data producer.

3.0.6 Skill Gaps

VS is interested in and currently pursuing a position with a skillset in both Bio Sciences and IT to better suit the needs of their scientists. Some of the scientists in VS have adopted knowledge of the information systems needed and are fulfilling a dual role of system owner/creator and scientist. While this is extremely helpful to the already stretched IT resources, it does detract from the scientific mission of the program.

Outreach between VS and IT in Ames is done by a monthly meeting with IT where scientists and IT exchange ideas and pain points about the environment to work together towards solutions. This effort has been very successful in better integrating IT into scientific computing. This has embedded a better understanding of each other's capabilities and needs.

VS has outlined that necessary skills needed for IT, Informatics and Data Scientists should be:

- Information technology resources to support the design, implementation, support and maintenance of network, computing, and data management resources that specialize in scientific computing.
- Informaticists skilled in designing data management strategies for scientific computing.
- Data scientists skilled in techniques such as data mining, machine learning, cluster analysis, and visualization along with mathematics or statistical analysis capabilities.

The Federal government has not yet created a job series for Informatics or Data Science making it difficult to align talent with the jobs needed.

3.0.7 VS Application and Database migration

Through interviews with VS, it was found a portion of the Veterinary Services Process Streamlining (VSPS), Veterinary Export Health Certification System (VEHCS), CVB Licensing, Serial Release & Testing (LSRTIS) and Phytosanitary Certificate Issuance & Tracking System (PCIT) applications are currently being migrated to a SalesForce Powered Certificates, Accreditations, Registrations, Permits, and Other Licenses (CARPOL) application which will reside in the SalesForce cloud.

Plant Protection and Quarantine (PPQ) Summary of Findings

1.0 Background

APHIS' Plant Protection and Quarantine (PPQ) program safeguards U.S. agriculture and natural resources against the entry, establishment, and spread of economically and environmentally significant pests, and facilitates the safe trade of agricultural products. PPQ accomplishes this through:

- **Plant Pest and Disease Programs -** Protecting Agriculture and the Environment from Invasive Plant Pests and Diseases: An invasive pest is a non-native species whose introduction into the country can cause damage to the economy, natural resources, or human health.
- **Plant Health Import Information -** Establishing Effective Regulations and Policies: By determining which plants and plant products can be imported—and which pose a high risk and should be excluded—the regulations and policies established by PPQ to protect the environment and U.S. agriculture.
- **Center for Plant Health Science and Technology -** Safeguarding Through Science: APHIS scientists monitor data from around the world and throughout the country to uncover pathways and develop strategies to both exclude pests before they arrive at our shores and to stop or limit their movement if they enter the country.
- **Phytosanitary Certificate Issuance and Tracking System -** Assisting U.S. Farmers and Exporters: APHIS assists American farmers and exporters by providing plant health inspection and certification for plants and plant products being shipped to foreign countries. Required by importing countries, these plant health certificates ensure that products are pest and disease free.

2.0 'As-Is' Environment

PPQ is spread out in many locations with very different needs. The first set of onsite interviews occurred at Beltsville, MD, building 580 where we met with the Center for Plant Health Science and Technology (CPHST) and The Plant Germplasm Quarantine Program (PGQP) lab managers.

Both CPHST and PGQP have the largest need for Big Data in PPQ which parallels what VS is doing on a smaller scale with genetic sequencing. Currently, these labs do not have any type of centralized or dedicated computational platform or any form of dedicated hardware for computational processing or storage on-site. Instead, a majority of the work is done on local workstations or decommissioned hardware that is air gapped from the network for running custom scripts for day-to-day tasks.

Additional locations were identified by Michael Stulberg who stated: "While PPQ is spread out with different needs, in Science and Technology there are other locations that are performing similar work to Building 580, albeit on a slightly smaller scale. Mission Lab in Texas is currently analyzing Illumina data (sequenced off-site) and is acquiring a MinION (portable, real-time biological analyses) to produce data on-site. The MinION generates in the ballpark of 100GB of raw data per sequencing run. There are no centralized computational platforms at any of the PPQ sites. Another lab, Otis in Massachusetts, will be getting into NGS sequencing in the near future. A lab being built in Sacramento, CA (still on the drawing board and has not yet been completely planned out) will also likely be analyzing NGS data. There is also a Ft. Collins lab separate from the building with VS that could get into NGS sequencing in the future."

While interviewing PPQ in Raleigh, NC, and Riverdale, MD, it was found that their environment was designed to support regulatory field work so the bulk of their Information Systems are located either in NITC's network or sponsoring educational institutions.

PPQ, beyond Building 580, currently has no big data needs though this may change in the future. The PGA dataset has the potential to grow into a large dataset that is interconnected with a lot of external federal systems for the purposes of long-term statistical models. However, at this time, the PGA Dataset is merely 26GB in size with a growth potential to 50GB over the next few years. The data resides in XML format and is currently held in a model on a PPQ workstation using a Microsoft SQL Database and Microsoft Power BI. PPQ has a strong desire to bring this dataset to a multi-agency and multi-dataset integrated platform that would allow for more powerful analytics in the import/export of animal and plant materials. However, NIS has large databases of sequences that they would like to share. Most are located within the B580 (Beltsville) area, however, some new hires will be located near the Smithsonian in Washington, DC.

2.1 Storage

CPHST and PGQP Labs are the largest producers of Big Data in the PPQ Program. Working with Next Generation Sequencers, they are on target to produce 30TB of raw and analytical data per year once at full capacity, possibly growing at a rate of 1TB per month. The labs are just now beginning to use the sequencers and produce roughly 100GB of raw data per sequence which turns into 300GB after undergoing analytics.

Local workstations, portable hard drives, and the enterprise storage environment are where a bulk of the data is kept for the labs. The 2TB hard drives that the labs were using were quickly filled up and new drives had to be procured. Mission Labs currently has data that could be stored as well and both the Mission and Otis Labs are expected to run into similar problems in the near future.

Mission Labs does very similar work in analytical sequencing that is similar to the Beltsville, MD Building 580 labs. It is expected to exceed its capacity far before Otis Labs. Upon interviewing the staff at Mission Labs, it was found that their current network connection is not adequate for downloading the large datasets that are produced by external sequencing that takes place in labs that they send samples to. When Mission Labs staff downloads the raw datasets for analytical work on their 20Mbps connection, the datasets of 5 – 10TB could take a week or more to finish. Instead of downloading the dataset, the USDA ARS colleagues co-located at Moore Air Base travel to their nearby University partners at the University of Texas Rio Grande

Valley where they have a 100Mbps Internet2 connection more capable of downloading the data in a much more reasonable time and transport it back to their labs on external portable media. Mission Lab staff are considering this method and will look into it in the near future.

Other PPQ databases such as the Global Pest & Disease Database (GPDD) and Spatial Analytic Framework for Advanced Risk Information Systems (SAFARIS) are housed in Raleigh, NC at the North Carolina State University (NCSU), which is not part of the APHIS network, but on a USDA-certified Server. CARPOL and ePermits are databases housed in NITC and are 800GB in total size.

CPHST and PGQP data retention requirements are an up to 10 years which should be considered when trending overall data growth.

2.2 Security

No special security requirements exist beyond the basic protections of APHIS data.

2.3 Data Sharing and Transfer

All data sharing is conducted via emailing small datasets that reside in files or transfer to portable hard drives and mailed out to the prospective location.

2.4 Scientific Computing

PPQ currently does not have any servers dedicated to scientific processing. Building 580 received a Linux Red Hat Enterprise Server similar in specification to the RHEL servers located in Ames, IA at VS. This server is not currently on the APHIS network and is pending evaluation prior to being joined with the network.

CPHST and PGQP Labs are interested in gaining access to the ARS SCINet environment to be able to access larger computational resources. Currently, no Internet2 connection exists at the Building 580 location and the nearest access is at the National Laboratory a few miles up the road where a 100Gbps Internet2 connection exists. A network extension would need to be completed to that PoP in order for Building 580 labs to have access to SCINet.

The closest Internet2 hub for Mission lab might be in either Houston (layer 1-3, 343 miles) or San Antonio (layer 1,235 miles) for at least a 10Gbps connection. The IT infrastructure in Mission, TX may need evaluation to get a minimum decent internet connection to SCINet.

The closest I2 hub for Otis Labs might be Brown University, 60 miles away.

The Ft. Collins location is roughly 6 miles from CSU campus and the Animal Science building is unable to take advantage of the high-speed SCINet connectivity without traveling to the campus.

2.5 Cloud

The only usage of cloud applications is in the form of the databases and web based applications that are hosted in CARPOL and ePermits at NITC in Kansas City, MO. Other applications do have a data feed into ePermits such as the International Trade Data Systems (ITDS), which takes Customs and Border Protection (CBP) data and electronically merges it with APHIS ePermits streamlining the import/export process of plants and animals as required in the Executive Order: Streamlining the Export/Import Process for America's Businesses (https://obamawhitehouse.archives.gov/the-press-office/2014/02/19/executive-order-streamlining-exportimport-process-america-s-businesses).

3.0 Miscellaneous Findings

3.0.1 Unmanned Aircraft Systems (UAS)

PPQ S&T is looking to use quadcopters (a type of UAS) equipped with high-resolution cameras UAS's in order to increase the capabilities ground survey teams to detect signs of Asian Longhorn Beetle infestation at the top of trees.

Further investigation into the use of multi-spectral cameras mounted to a UAS is also being reviewed to track invasive species of plants or pests over large areas in a small amount of time thus increasing the responsiveness to any emerging crisis. Using UAS with devices mounted to release sterile male insects to control pest populations is also being contemplated.

The initial exploration of the technology would only generate telemetry data, photos and video locally to the sensor mounted on the UAS and to a field laptop that would be used in controlling of the UAS's. This data would eventually be brought back for analytical work and storage on the Enterprise Network as it does not require any large computational resources.

3.0.2 Open Source

CPHST and PGQP are eager to start using the same open source products listed below. Several of the lab directors from CPHST and PGQP have been in contact with VS about the process to levy open source technology.

Unfortunately, the open source packages they were trying to get integrated into their lab environment were grandfathered in at VS and were not able to be installed for PPQ. The below open source packages are available on SCINet and would fulfill the immediate needs of the labs:

OPEN SOURCE SCIENTIFIC APPLICATIONS			
BWA	ABySS	Shell cmds	RAxML
GATK	Kraken	Git	t-coffee
Picard	kSNP	Perl	LaTeX
Samtools	Bamtools	Python	
IGV	Bbmap	R	
SPAdes	Newick utils	Local BLAST	

PPQ and Mission Labs would also benefit from being able to rapidly test software or software packages using a sandbox environment that was recommended for VS.

Biotechnology Regulatory Services (BRS) Summary of Findings

1.0 Background

In order to protect plant health, Biotechnology Regulatory Services (BRS) implements APHIS regulations for certain genetically-engineered (GE) organisms that may pose a risk to plant health. APHIS coordinates these responsibilities along with the other designated federal agencies as part of the Federal Coordinated Framework for the Regulation of Biotechnology.

- **Regulations**
  Established as a formal policy in 1986, the Coordinated Framework for Regulation of Biotechnology describes the Federal system for evaluating products developed using modern biotechnology.
- **Permits, Notifications, and Petitions**
  APHIS regulates the introduction (importation, interstate movement, or environmental release) of certain genetically-engineered (GE) organisms. All regulated introductions of GE organisms must be authorized by APHIS under either its permitting or notification procedures.
- **Compliance and Inspections**
  BRS has a comprehensive system to help ensure that biotechnology organizations are maintaining compliance with APHIS' biotechnology regulations.
- **BQMS Program**
  The Biotechnology Quality Management Support (BQMS) Program helps biotechnology researchers and organizations analyze the critical control points within their management systems to help them better maintain compliance with APHIS regulations.

2.0 'As-Is' Environment

BRS is not a research organization but a regulating body which consists of Regulatory Operations, Biotech Risk Analysis and Resource Management. BRS currently conducts 750 – 850 inspections per year where it generates an 8- to 10-page report that is fed into the ePermits and COGNOS system hosted in Kansas City, MO at NITC.

2.1 Storage

Enterprise Storage is the primary medium that is used for storing documents and any artifacts. The entire size of this environment is estimated to be less than 1TB in total.

While all the permitting information is kept in ePermits at NITC, some of the information is copied externally to the Information Systems for Biotechnology at Virginia Tech (http://www.isb.vt.edu/) which is also accessible via the BRS Permits site (https://www.aphis.usda.gov/aphis/ourfocus/biotechnology/permits-notifications-petitions/sa_permits/ct_status). The total size of this Information System is around 800GB on an Oracle Database located at NITC. This database grows less than 1GB per year.

GIS information is stored locally on laptops or workstations. Other than file shares on the Enterprise Storage platform, there is no formal storage or retention policy.

2.2 Security

Due to the nature of the reports, they have a need to remain confidential and are stored in the ePermits system. Either Cognos reports or ePermits information can be shared externally by queries of the system or downloaded as a PDF file and emailed to the recipient.

2.3 Data Sharing and Transfer

To conduct field inspections, a fillable PDF is exported from ePermits which the inspector loads onto their laptop and it is filled in during the inspection process. Unfortunately, there is no way to upload the fillable PDF to ePermits for data interchange. Instead, the inspector has to manually type in the findings and attach any supporting documents.

GIS Information is typically obtained from outside sources from a State or Federal database (Fish & Wildlife Service). Finding GIS data can be challenging and some research has to be done for each dataset pulled. Once the data is sourced, it may need substantial reworking prior to it being used in BRS products.

2.4 Scientific Computing

BRS has neither a scientific computing environment nor a need to access one. All of the computing is done via web accessible applications such as ePermits and COGNOS. Workstations and enterprise storage are used to carry out day-to-day business while laptops are used in the field for inspections.

2.5 Cloud

BRS is using ARCGIS from Esri and plans to migrate to the cloud version of the software as part of the overall effort by APHIS for all of its programs.

Systems which are housed at NITC in Kansas City, MO could be construed as Cloud-based platforms such as ePermits and COGNOS.

The ePermits Information System is currently being replaced with eFile, a SalesForce empowered system located directly in the SalesForce Cloud as a SAAS offering.

3.0 Miscellaneous Findings

BRS has cited the need for a real-time set of species layers that are sourced from Department of Fish & Wildlife Services. The current method of sourcing is difficult and some of the information is found to not be as up to date as BRS analysts would like to run accurate models. Metadata sources aren't being tracked and could lead to inconsistent datasets if standard or centralized sources aren't being used.

Wildlife Services (WS) Summary of Findings

1 Background

The mission of USDA APHIS Wildlife Services (WS) is to provide Federal leadership and expertise to resolve wildlife conflicts to allow people and wildlife to coexist. WS conducts program delivery, research, and other activities through its Regional and State Offices, the National Wildlife Research Center (NWRC) and its Field Stations, as well as through its National Programs.

Program biologists apply the integrated wildlife damage management approach to provide technical assistance and direct management operations in response to requests for assistance. WS NWRC research scientists are dedicated to the development of wildlife damage management methods.

The program's efforts help people resolve wildlife damage to a wide variety of resources and to reduce threats to human health and safety. Funding for the WS Program is a combination of federal appropriations and cooperator-provided funds.

Wildlife Services conducts its activities pursuant to Memoranda of Understanding, other agreements, and legal authorities, and conducts environmental review processes to comply with the National Environmental Policy Act (NEPA). WS develops Annual Program Data Reports to provide the public with information about its wildlife damage management activities.

It was found that WS has many databases that need to be centralized:

- Management Information System (MIS)
- National Feral Swine Database
- National Rabies Database
- Wildlife Scientific Data Archives
- Wildlife Tissue Archives

All of these database systems in WS could benefit from Big Data, analytics, and interconnections on a centralized platform that WS had direct control over.

The reduction in on-premise servers has troubled the stakeholders in WS. Their field sites are losing precious computational assets while the network connectivity on the APHIS UTN for these sites is not commensurate with the growth needs to access the proposed centralized or cloud platforms.

An interview with the APHIS CTO has shown that sites are being upgraded from legacy T1 connections to 5 – 20Mbps Ethernet connections. This will go far to start solving the slow networks, but WS is skeptical that these small upgrades will be enough to ensure a usable platform at the APHIS Enterprise Hub Sites (Riverdale and Ft. Collins) or in Microsoft Azure.

To complicate matters, the bulk of WS leadership is not collocated on the Ft. Collins Campus, but 6 miles away at the CSU Foothills Campus, making it impossible for that site to benefit from the centralization of systems located within the APHIS Ft. Collins Site.

While greater connectivity has been offered to WS and their field sites, the issue is budget based in that IT needs are exceeding what is allocated for the budget.

2.0 'As-Is' Environment

WS is divided into two operational branches: the eastern region is based in Raleigh, NC, and the Western Region in Ft. Collins, CO, in addition to eight field sites. Further, the National Wildlife Research Center NWRC (the research division of Wildlife Services) is also headquartered in Ft. Collins, CO. The NWRC has eight field stations with a majority of the work effort being done in Ft. Collins, CO.

The NWRC field offices are connected to Ft. Collins via a T1 connection that is slowly being upgraded to Ethernet. Issues arise in the field offices when any amount of information is copied across the network, which causes network outages until the transfers are completed.

Some of the field offices have servers, but no technical information has been gathered on them to glean their computational power. The old servers are being replaced due to the requirement of APHIS to eliminate most on-premise servers. However, what options (e.g., cloud-based, hybrid servers, etc.) available for replacement are still unknown.

2.1 Storage

WS has between 29 - 32TB of total data broken down by the following locations on a Windows 2012 Storage Server:

- Florida – 14TB
- Mississippi – 8TB
- Utah – 5TB
- Hawaii – 2- 5TB

While these locations do not comprise the entirety of the WS data, they are the largest contributors and should be taken into consideration as APHIS looks to consolidate physical systems at either a hub site (Riverdale or Ft. Collins) or the Microsoft Azure Cloud.

This data footprint does not include external storage media that has been used to backup or transport the data between sites. Overall data growth is expected to be 1TB a year on the average.

All other storage needs are met by the Enterprise Storage Platform, but the reduction in local footprint at the field sites is still a major WS concern.

2.2 Security

Some of the data that WS produces falls under CIPSEA, making distribution and safeguards of that data a higher priority than other data types in APHIS.

2.3 Data Sharing and Transfer

Most of the WS sites are self-contained and require minimal data sharing, with most of the data on-site for analytical work. The exception to this is national programs (e.g. Feral Swine and Rabies), which have a need for centralized and large data storage. If the data needs to be transferred, the site typically uses the APHIS network connection to transfer the datasets between sites. This does cause issues with the network on remote sites connected via a T1 line rendering them almost inoperable while the data transfer occurs. Transferring data between Hawaii or Guam is the most challenging as they have the slowest network connectivity and lack of resources. Data from these sites may be shipped on portable storage media between sites.

WS uses measurement devices in the field that transmit their data back to WS analysts in the form of emails. Each file contains the raw data that is aggregated into analytical reports.

2.4 Scientific Computing

WS has a lab in Ft. Collins that contains 11 workstations as a local computational resource to run models and perform analytics. Out of the 11 available workstations, only 8 can be used at one time in the lab due to building power constraints. The workstations are typically removed from the APHIS Enterprise Network during processing so that the Enterprise Updates do not interrupt long-running models or analytics that may be occurring on them.

An interruption in service denotes that there is a need to segregate scientific computing from the Enterprise Network.

A server was added to the National Wildlife Research Center (NWRC) lab 2 – 3 years ago for the purposes of Big Data processing. The lab manager was not available for getting the specifications of this system during the interview process.

2.5 Cloud

WS is behind on using ArcGIS from Esri and plans to utilize the cloud version of the software as part of the overall effort by APHIS for all of its programs. The cloud option for ArcGIS has not been available to WS users. Further, given the nature of the work conducted by many WS programs, the cloud version will not provide any advantages. As stated above under the VS section, WS GIS users are distributed in offices (many of which are remote) throughout the USA (and the world), so although cloud-based solutions might look appealing, the current state of connectivity and computer performance (without major upgrades) will likely make this an unreachable/impractical goal.

While the utilization of Esri cloud is being encouraged by APHIS Enterprise IT, there are analytics that consist of so many data points that they are unusable in the Esri cloud. Such workloads should be considered for large on-site workstations which do not meet the DCOI reduction requirements to ensure that computational resources for large GIS datasets are achievable at field sites.

3.0 Miscellaneous Findings

WS pulls data from each of the states for its work products. A server is located in each state (40 in total) and is connected to the APHIS network. The states are not unable to share their information between other states, nor are the WS users allowed to share this data, thus leading to stove pipes in sourcing information. This is a tedious and time-consuming process to aggregate the information from the states for WS work products. However, for many of the WS programs, they are NOT ALLOWED to share data based on the wishes of their cooperators.

WS is not confident in their local backups. As a side effect, analysts tend to horde information onto external hard drives that aren't backed up or part of the overall COOP strategy, which could lead to the destruction of long-running datasets in the event of a failure or accident.

On February 22, 2013 (NWRC is looking at this very closely because it affects them greatly), former President Obama released an executive order that states that the public has a right to transparency of research data that was funded by public money. This memorandum (https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf) could have far reaching effects on how research data is stored and how it has to be shared with the public.

In being forward-looking for the Presidential mandate, WS is very concerned about the way that data can be searched on their public-facing internet pages. The current Enterprise Indexing system leaves much to be desired as queries tend to have nonsensical returns that confuse the customer who is trying to source the information. WS implemented a search solution through an outside provider who compiled the indexes for WS and allowed for better search results to be returned with a high degree of success and positive customer feedback. Since the provider is not FEDRAMP certified, WS had to take their efforts offline with the current index provider for their meta-data and revert back to the Enterprise IT provided solution. While this solution was ideal for the time it was being used, WS has since started to look towards the U.S. Forestry Service (USFS), who has implemented a successful public information portal of their tax-funded scientific research. WS has since pursued the possibility of a joint portal since USFS has had success in meeting the Presidential mandate.

WS is currently maintaining their archival solutions by manually hashing all of the files that they ingest into their archive using an open source program called "droid". This is a time-consuming process and WS wishes to automate or incorporate these data integrity checks into their future archival platform.

Animal Care (AC) Summary of Findings

1.0 Background

USDA Animal Care is staffed by experts on animal care and husbandry who help lead the way in determining standards of humane care and treatment for animals. Animal Care works with states, regulated industries, animal advocacy groups and other non-governmental organizations to ensure the welfare of millions of animals.

- **Animal Welfare Act**
  The Animal Welfare Act and its associated regulations require that federally established standards of care and treatment be provided for certain warm-blooded animals bred for commercial sale, used in research, transported commercially or exhibited to the public.
- **Horse Protection Act**
  The Horse Protection Act and its associated regulations seek to put an end to soring (a procedure in which horses are subjected to chemical and/or mechanical irritants in order to enhance their gait) by preventing sored horses from participating in exhibitions, shows, sales, or auctions.
- **Center for Animal Welfare**
  The Center for Animal Welfare collaborates with other animal welfare entities to play a central role in USDA's efforts to build partnerships domestically and internationally, improve regulatory practices, and reach beyond USDA's traditional enforcement role to develop outreach, training and educational resources.
- **Animal Care Emergency Programs**
  USDA Animal Care's emergency response component provides national leadership on the safety and well-being of pets during disasters – with the understanding that supporting animal safety during emergencies is a significant factor in ensuring the well-being of pet owners.

2.0 'As-Is' Environment

AC consists of 115 inspectors that conduct on the average 11,000 assessments per year and enter their results in the Animal Care Information System (ACIS3) hosted in Kansas City, MO at NITC.

2.1 Storage

The entirety of the AC permitting data is kept in ACIS3 and totals about 800GB in size. This information system stores just the raw data and any supporting documents are kept in some sort of file share outside of ACIS3. When a report is brought up and any pictures or video clips need to be reviewed, they have to pull the permit or report in a special way that's not intuitive to see the supporting documents.

AC does conduct one off studies in the field but contains all of the data on a laptop of the scientist doing the study and backs the information up off the laptop onto 1TB portable storage media. This storage media is kept in the lab unless traveling, and then it is kept with the laptop.

If the laptop or portable storage have any data loss or are destroyed, then valuable scientific data may be lost.

AC enacts a retention policy of 3 years for ACIS3 data in general and 10 years for any permit that is or has been in legal dispute. However, currently due to issues with the ACIS platform, no data is being purged.

2.2 Security

No special security requirements were found to exist beyond the basic protections of APHIS data.

2.3 Data Sharing and Transfer

AC being regulatory in nature shared permit information only via ACIS3 or extracted PDF reports. If there is a legal dispute in the permitting process, then a PDF file will be generated and shared with the appropriate regulating bodies as evidence.

AC conducts one off field studies which are mostly analytical in nature. The product of these studies is a PDF file which is released through the APHIS AC website or to external sources for collaboration via email.

AC permitting accounts for 2/3 of all FOIA queries to APHIS.

2.4 Scientific Computing

AC does not have any scientific computing environment. All source data and analytics are collected on individual workstations or laptops.

AC is conducting outreach programs to educate its target audience on subjects where they are seeing the greatest challenges in the permitting process. AC would like to trend whether violations go up or down in proximity to these events. The current ACIS3 does not account for them being able to cross-reference attendees with violations in the permitting process to obtain this information. AC is seeking a quantitative approach to better the permitting process and lower violations in animal environments.

2.5 Cloud

The only cloud application or system AC is accessing is ACIS3 located at NITC. There were no other cloud applications or products that ACD would need in the future.

3.0 Miscellaneous Findings

AC has noted that when it queries information in ACIS3, the results are not consistent. Issues have been found after the closeout of the fiscal year when reporting reflects 8 – 12% deviation in the numbers for multiple inspections completed. Some of the issues driving this result from how data is classified. If an inspection is edited or updated, it could count as a new report. If reports aren't closed or are still pending data, they may completely fall off the reporting process.

Through interviews with AC, it was found that the ACIS3 application is currently being migrated to the SalesForce powered CARPOL application.

CNSS Summary

CNSS has conducted on-site interviews in Ames, IA, Ft. Collins, CO, Beltsville, MD and Riverdale, MD and numerous field sites via tele-conference over the course of this project. Phone interviews were conducted for Raleigh, NC and all program personnel that were unable to attend in person, as well as a live demonstration of the SAFARIS system conducted by teleconference. CNSS also conducted interviews at the request of PPQ with their Mission and Otis Labs locations. CNSS conducted interviews with additional staff from WS at the request of Brian Washburn and Larry Clark:

- NWRC Scientific Data Archiving / Information Services
- MIS (WS' Management Information System)
- NWRC Genetics
- Feral Swine (Research) Program
- National Rabies Management Program

CNSS found that the largest users of Scientific Computation Environments were VS NVSL (Ames, IA and Plum Island, NY), CVB (Ames, IA) and CEAH (Ft. Collins, CO). PPQ CPHST and PGQP Labs located in Building 580 in Beltsville, MD are the largest consumers in the PPQ program, although Mission and Otis Labs have the propensity to become massive data producers once their platforms get online and to capacity. Finally, WS has a single lab in Ft. Collins, CO that has a large computational need.

Taking into account past and present trends, all sites described are on target to need a dedicated scientific environment to meet data growth projected to be in the Petabytes. A short-term solution to address current gaps will need to be identified since creating a dedicated scientific environment will be a long-term solution.

Each program has a need for similar tool sets as they use identical next generation sequencing technologies, but modeling has some unique tools that are not open source and require greater processing power. Through the interview process, we also found that VS has open source tools that PPQ labs were not able to take advantage of due to security constraints preventing their installation and use on the APHIS corporate network.

Exchanging information between sites, programs, and even external collaborators has proven to be a challenge for APHIS as a whole. Due to the lack of storage resources or connectivity, the programs have resorted to using portable storage media and mailing the devices. This was one of the main points of emphasis from several of the interviews.

Connectivity at the field sites is the number one issue next to lack of computational platforms on-site. While the T1 lines are upgrading to 5 – 20 Mbps Ethernet line, concerns around the bandwidth being inadequate as the DCOI consolidation will push over 171 servers from sites under 20 personnel to the hub sites (Riverdale and Ft. Collins) or Microsoft Azure.

Most sites have local computational resources in the form of workstations or reclaimed IT assists. Due to the ad-hoc growth, some locations cannot power all of their workstations in the ad-hoc labs due to building power constraints. Some of the scientists are removing enterprise systems from the network to stay online without interruption from updates while their computations complete.

Some scientists have found that these ad-hoc environments impede the timely processing of the data and have turned to IT for the more centralized computational resource. These have been provided in the form of centralized servers in VS and PPQ running Red Hat Enterprise Linux. Most of the scientists do not have experience in a command line environment and are more comfortable in a GUI based environment leading to lack of adoption of the platform provided. Modeling is also in need of an ad-hoc environment for timely processing. Working in a command line environment is not a huge barrier for them but rather the lack of access to resources.

The main CEAH areas of need are:

- A short-term solution for processing speed (e.g. a local cluster) while the longer-term solution (e.g. SCINet etc.) is worked out, as this is likely to take years. They need faster processing now.
- A short-term solution for centralized data storage with automatic backup
- Better options for moving large amounts of data internally and among external collaborators
- Human resource needs for people with the right IT/scientific skill sets.

Despite the challenges that the programs in APHIS have faced, they continue to find new and novel ways to reclaim IT assets or cleverly relocate data getting around the lack of storage available. Scientists and IT have been proactive in VS by forming working groups to help isolate needs and work cooperatively together towards solutions within budgetary constraints.

APPENDIX A

Challenges

Below is a list of challenges by APHIS programs that have been identified throughout the Summary of Findings. The intent of this redundant list is to extract the needs in a single list that is more concise and easy to read.

The summary below is presented chronologically vs. by program priority from the overall findings in this document.

Veterinary Services (VS)

- Lack of centralization of all scientific data
- Storing some of its scientific data on the Enterprise Network
- Enterprise IT to require the scientists to remove data from the enterprise systems in response to resource issues
- Scientists store scientific data in their Enterprise Share
- Lack of storage resources are forcing scientists on to memory sticks, hard drives or other network shares
- Storage data is growing at a rate of 235% per year
- Storage medium used for the storage of Scientific data is near end of life
- Externally curated datasets have flaws in them that have to be fixed prior to the scientists or analysts being able to utilize them
- Metadata is not indexed to prioritize aging datasets into proper storage tiers
- Datasets with PII, which require higher protections, cannot be moved to external systems without the proper safeguards or accreditation
- Pseudo air-gapped non-complaint scientific instrumentation through multi-homed Windows systems as a work around to satisfy security
- Data transfer may have to be done by copying data to portable media and mailing them between sites
- Scientific compute resources in Ames, IA are not clustered and there is no real scheduling other than manual coordination for priority or access to them
- Use of FTP to move data between locations has security implications
- APHIS routes all traffic bound for SCINet through the TIC and subsequently an AWS environment before landing in SCINet creating a slow bottleneck that takes long periods of time for basic data transfer
- Scientists are forced to use ad-hoc (Workstations & Laptops) resources for their computational needs as a work around for lack of a centralized Scientific Computing environment
- Current Cloud initiatives for Azure does not answer the need for a scientific computing environment and caters more towards Enterprise Data

- Connectivity between Plum Island, NY, and Ames, IA is inadequate for the transfer of large datasets, manually shipping of data occurs as a work around
- Use of open source tools for evaluation takes months to years in some cases
- No formal agreements exist between ARS and APHIS for use of the SCINet environment
- Skill gaps exist between scientists and IT, that hamper scientific computing efforts
- Positions for Data Scientists and Informaticists have not yet been coded by the Federal Government leading to a challenge filling these positions in VS that need them
- A short-term solution for processing speed (e.g. a local cluster) while the longer-term solution (e.g. SCINet, etc.) is worked out, as this is likely to take years. They need faster processing now.
- A short-term solution for centralized data storage with automatic backup
- Better options for moving large amounts of data internally and among external collaborators
- Human resource needs for people with the right IT/scientific skill sets.

Plant Protection and Quarantine (PPQ)

- Scientists are not afforded the same types of open source tools that other programs are authorized for use
- Rapidly growing datasets are exceeding local storage and ad-hoc methods are now being used such as portable storage media
- Scientists are forced to use out of date or discarded hardware as their scientific computing environment
- Contemplating utilizing of UAS to meet the needs of field teams; has not yet been decided how data will be consolidated or integrated into the scientific analytical environment
- Data ingest in ePermits is a manual process with no automated data interchange that proves to be tedious and time-consuming

Biotechnology Regulatory Services (BRS)

- Data resides on individual workstations or portable storage mediums that aren't part of the enterprise backup strategy
- Has to manually enter data from ePermit inspections vs. automated data ingestion from the fillable PDF generated from ePermits at the beginning of the data gathering process
- Information gathered from States or Federal Fish & Wildlife databases are not accurate leading to datasets having to be fixed prior to being used in analytical products
- Lack of real-time species layers leads to inferior analytical products

Wildlife Services (WS)

- Datasets are compartmentalized from state to state on local servers and are not able to be shared leading to difficulties in curating data from more than a single source

- Field offices are connected via T1 lines which often are saturated when data interchange occurs between sites
- Field offices contained servers, but have not been inventoried and the replacement method has not been figured out
- Data stored on external media are not included with enterprise COOP plan and may be prone to irreplaceable data loss if they fail
- Programs like Feral Swine & Rabies have a need for centralized storage of large datasets
- Network connections to transfer data between sites is inadequate for their current needs
- Power to the Ft. Collins lab is inadequate to run all 11 workstations concurrently
- Workstations are removed from the APHIS network to facilitate processing without interruption from enterprise updates
- Cloud version of ArcGIS has not been available for WS users to experiment with
- States are not able to exchange information leading to stove pipes in information which has to be tediously collected and aggregated by WS for analysis in its products
- WS is not confident in their local backups and are unsure if they are included in the overall COOP strategy
- Past issues with Enterprise IT policies led to weeks on a project to consolidate the SharePoint Server which was heavily customized by a contractor
    - After the policy issues were fixed, the migration went smoothly
    - Took 6 weeks for remediation of the policy issues to get the WS contractor access to the SharePoint instance ITE was hosting, costing extra time and money on the contract
- WS CSU Foothills Campus is miles from the APHIS Ft. Collins Campus
    - Unable to take advantage of optimizations to this campus
- WS desires greater connectivity, but in select fields sites this is not possible due to budgetary issues
    - WS is concerned that consolidation and lack of bandwidth will cause issues to its mission
- Manual processes around ensuring archival integrity

Animal Care (AC)

- ACIS3 reports containing artifacts reside on an external system and confusion arises on how to access the artifacts of the report
- Accuracy of the overall permits in ACIS3 can change between queries leading to an inaccuracy of the overall sum or permit done in a fiscal year by 8 - 12% leading to possible inaccuracies in reporting
- Needs to have a way to quantify the seminars and outreach programs for impact on overall violations found during inspections

- Stores all data on local workstations or portable media not part of the enterprise backup strategy